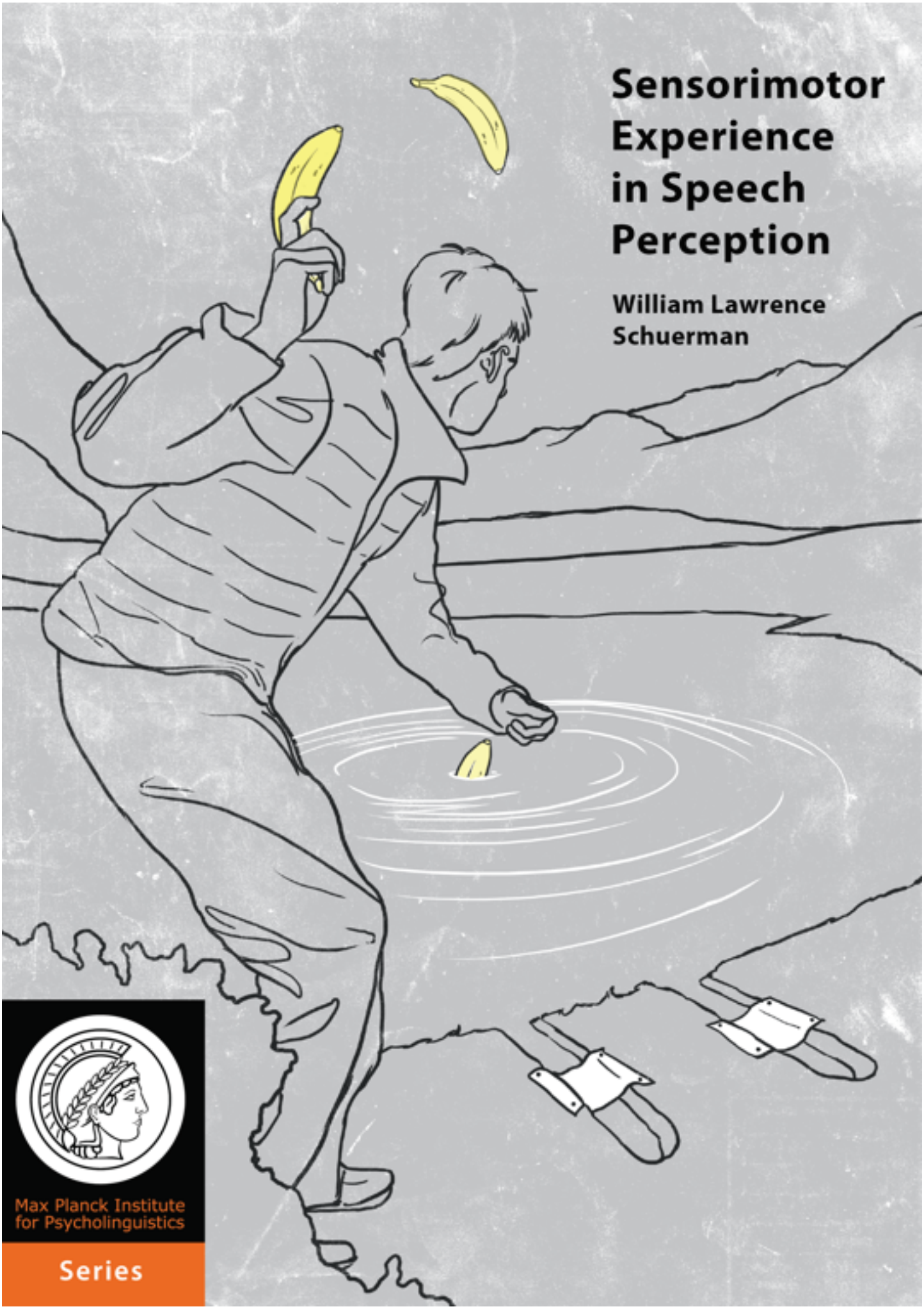
Sensorimotor Experience in Speech Perception

William Lawrence Schuerman



Max Planck Institute for Psycholinguistics

Series

Sensorimotor Experience in Speech Perception

ISBN: 978-90-76203-82-9
Cover photo: Kees Peerdeman, “Sounds bananas”, referring to the metaphor by Albert S. Bregman, likening auditory perception to observing the ripples made by objects moving on a lake.
Printed and bound by Ipskamp Drukkers b.v.

Sensorimotor Experience in Speech Perception

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van de decanen
in het openbaar te verdedigen op
maandag 27 maart 2017
om 12.30 uur precies

door
William Lawrence Schuerman
geboren op 28 september 1984
te Sonoma, Verenigde Staten

Promotoren:

Prof. dr. J. M. McQueen

Prof. dr. A. Meyer

Manuscriptcommissie:

Prof. dr. M. Ernestus

Prof. dr. H. Bekkering

Dr. P. Adank (University College London, Verenigd Koninkrijk)

This research was supported by the The Max Planck Society for the Advancement of Science, Munich, Germany.

Sensorimotor Experience in Speech Perception

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Monday, March 27, 2017
at 12:30 hours

by
William Lawrence Schuerman
born in Sonoma, USA
on September 28,1984

Supervisors:

Prof. dr. J. M. McQueen

Prof. dr. A. Meyer

Doctoral Thesis Committee:

Prof. dr. M. Ernestus

Prof. dr. H. Bekkering

Dr. P. Adank (University College London, United Kingdom)

This research was supported by the The Max Planck Society for the Advancement of Science, Munich, Germany.

Acknowledgements

In taking this opportunity to acknowledge the many people who made my doctoral research possible, it is regretfully inevitable that I will fail to name many who deserve to be recognized. For this my deepest apologies. This thesis would never have been completed were it not for my family, friends, colleagues, and supervisors. I say this not only out of gratitude, but in recognition of the extensive social support structure behind every line of text in this book. Thank you all for helping me do this.

To begin the particulars, I wish to thank my fellow PoL group members. You formed the cornerstone of my MPI experience, and the atmosphere of shared camaraderie was one of the greatest parts of coming to work. There are way too many of you to thank, so I'll just say thank you to you all (like a group hug).

Carrying out research required constantly grappling with new ideas and techniques. Had I not been able to turn to colleagues who graciously shared their knowledge, I could have easily fallen far behind. In particular, thank you to Seán Roberts and Mitty Casillas-Tice, for introducing me to trees, forests, and many other analytic techniques. Thank you to Linda Drijvers and Valeria Mongelli for sharing your expertise, scripts, and time. Also thanks to numerous other people in all the departments who at one time or another I went to for help. The MPI's greatest resource was your support and generosity.

Special thanks are due to my incredible officemates, paranymphs, and friends, Johanne Tromp and Merel Maslowski. Working, commiserating, rejoicing, and drinking lots and lots of coffee together created a gezellig little place to go every day. Johanne, yeah, I mean, come on! You are an awesome officemate, a great friend, and a very good dancer. Merel, thank you for sharing high fives, endless cups of coffee, and being exactly the way you are. I really lucked out sharing 363 with you both.

One of the greatest things about being at the MPI was being able to participate in a wide range of discussion groups and collaborations that broadened my perspective on a wide range of topics. Thank you to all the members of the Radboud University Sound Learning group, for the excellent feedback on my experiments and sharing your work with me. Similarly, thank you to the Radboud University SPRAC group for excellent suggestions and interpretations. I want to sincerely thank the MPI conversation analysis meetings (and in particular the SiN group) for welcoming me in, teaching me the ropes of CA (many thanks to Elliott for guiding my education), and letting me take part in a really fun project. It. Is. The. Best. Period. Similarly, deep thanks to John Houde, Sri Nagarajan, and the members of the UCSF Speech Neuroscience Lab for letting me invite myself over to collaborate and learn their sound altering secrets.

To Prof. Mirjam Ernestus, Prof. Harold Bekkering, and Dr. Patti Adank, thank you for giving my thesis thoughtful consideration and provide such incisive remarks.

To the TG staff, in particular Alex Dukers, Ronald Fischer, and Ad Verbunt, thanks for your patience and ever-present assistance. Even when it started raining in the experiment rooms, you were all ready to help figure out what was going on. To the library staff, Karin Kastens and Meggie Uijen, thank you so much for finding numerous odd articles, getting through paywalls, and obtaining obscure books.

This next part is a bit hard to write, but I think it's important. Mid-way through my PhD I had a personal crisis and did not know if I could (or should) continue. Looking back now, I am so very glad that I saw it through to the end. To all of you, I am immensely grateful. I would like to specifically thank Marjoleine Sloos, Katie Drager, Nicolai Pharao, Erez Levon, Cynthia Blanco, Shiri Lev-Ari, Richard Todd, and all those at the Bias in Auditory Perception conference. It may seem a bit strange to call a small conference on phonetics as being 'life-changing', but that is exactly what it was. Marjoleine, thank you so much for taking the time to talk with me, and for showing me that it's okay to ask for breathing room when it is desperately needed.

To Johanne, I couldn't have asked for a better office-mate and sympathetic ear during that hard time. To Elliott, I hope you know how important your friendship and support was. To Huub, I hope you know that you helped me work through a lot of things just by sharing your time with me, and I am very glad to have become friends with you. To Antje and James, thank you for showing such support for my emotional well-being, and such willingness to help me through that long rough patch. Your kindness meant a great deal.

To Elliott, thank you for adventure times, for going to distant lands together, for being critical, for introducing me to new ideas, for listening and talking and just being one of the coolest people I have ever had the honor of being friends with. Maybe the real treasure was the journey. Maybe the real thesis is friendship.

Social life is just as important to surviving a Ph.D. as educational support. Sunday brunch crew, there was never a week so bad that a few beers, waffles, and hours spent with close friends couldn't make all better. Lisa, René, Rick, Ellen, David, the PHD band, we had a great run, thanks for making my last few months filled with so much music. My housemates at Le Fab, for turning dishes into dancing, birthdays into all-day brunches, living rooms into karaoke dens, a house into a home and friends into a family, thank you all so much. The cover art was graciously illustrated by Kees Peerdeman, thanks champ! Bart and Louise, for just being great people to cook, protest, and volunteer with.

To my supervisors, Antje Meyer and James McQueen. You allowed me to work independently, yet were always responsive and ready to give feedback and advice when needed. James, thank you for helping me delve deeper into phonetics and for offering great advice and feedback when I struggled with writing or analyses. Antje, thank you for always having an open door, for replying to emails with astonishing speed, and for providing critical feedback while always being encouraging.

Lastly, thank you to my family, for dealing with me running around the world, for encouraging my interests, and for always just wanting me to be happy. You even tried being participants in one of my weird experiments (sorry I put you through that... it may happen again). Love you lots.

Contents

List of Figures

List of Tables

Dedicated to my teachers, advisers, friends, and family, for all their encouragement, support, and caring.

Chapter 1

Introduction

1.1 Action and perception

"Why is the earth round? Why isn't it square or any other shape?" (Sagan et al., 1980). This may seem like a ridiculous question, but as Carl Sagan argued, these are exactly the type of questions we should be encouraging children, and each other, to ask. Why is the world one way and not another? Basic questions such as these spur us to examine our assumptions about the natural world, leading to new discoveries.

A similar basic question one might ask about humans (and many other animals) is why do we have senses. What is the need for vision, touch, or audition? Trees get along fine without eyes, ears, or noses, why should we need them? The answer seems to be that there is a fundamental difference between forms of life that have highly developed sensory organs and those that do not: Action. In order to move about in our environment in ways that benefit us as organisms, it is crucial to be able to gather information about that environment. Take the sea squirt; in its larval form, the sea squirt has a simple brain-like ganglion with which it can receive sensory information from its environment. However, once it attaches itself to a rock and ends the mobile period of its life cycle, this brain is mostly digested, suggesting that it is no longer needed once the organism becomes stationary (Llinás, 2001). Thus, the need to extract information from the environment appears to be predicated on the need to move through and act in that environment.

Rather than passively putting us in position to receive sensory information, interactions between action and sensation have been argued to constitute our perception of the world. In the context of vision, von Helmholtz argued that "when we perceive before us the objects distributed in space, this perception is the acknowledgement of a law-like connection between our movements and the therewith occurring sensations" (von Helmholtz, 1977, p. 138). In vision, sensorimotor experience refers to the experience of a shifting visual display that changes predictably as a result of self-generated actions, such as eye and body movements. For example, if we look at a car and it increases in size, this could be because we are walking towards it and the sensory image of the car comes to take up more of our visual field. Alernatively, if we are not moving, we perceive that

either the car is growing (not very likely) or that it is coming closer (more likely). Under this interpretation, perception is not a sensory but a *sensorimotoric* phenomenon, and *sensorimotor experience* refers to repeated exposure to these action-contingent changes in sensation.

One excellent example of a "lawlike connection" is the perception of "roughness". How can we know if something is rough? Roughness is a pattern of sensation created by physical contact as well as motion across a surface (O'Regan and Noë, 2001). Even when we are not actively touching an object, our accumulated knowledge about which visual properties co-occur with certain actions, i.e., our knowledge of sensorimotor contingencies, may enable us to discern tactile attributes of an object, such as roughness, smoothness, or sponginess, just by looking at it. In this way, sensorimotor experiences can be argued to actively shape our perceptions of the world.

Moving around in the world squishing various objects is not the only way to generate sensorimotor contingencies. For example, in addition to being perceivers of sound, we are also producers of sound. There exist lawlike connections between our actions and the auditory sensations that these actions produce. We clap our hands together hard and a much louder sound is produced than when we clap softly. This auditory feedback from our own actions, i.e., auditory sensorimotor experience, may alter how we perceive sounds generated by others (e.g., Paulus et al., 2012). With regard to speech in particular, our perception of speech may be shaped by repeatedly experiencing, during speech production, the sensorimotor contingencies between motoric movements of speech articulators (e.g., lips, tongue, larynx) and the sensory outcomes of these movements. This thesis examines to what extent sensorimotor experience acquired during speech production may influence speech perception.

1.1.1 Sensorimotor control theory and speech production

Sensory feedback is extremely important for speech production. Delay auditory feedback by a tenth of a second or so and the task of producing fluent speech becomes extremely difficult (Chase et al., 1959; Yates, 1963). The ability to monitor our speech during production enables us to rapidly detect and correct errors (Levelt, 1989). Observations about the patterning of speech errors and the speed of corrections led psycholinguistics to suggest two routes for error detection, one internal and one external (Levelt, 1983; Nozari et al., 2011). Just as the transmission of sound moves at a specific speed through a medium, the transmission of auditory information from our outer ear to the auditory cortex takes time. The speed with which speakers are able to correct speech errors, even within the breadth of a single segment, challenges the notion that we correct errors

by passively monitoring the sensory outcomes of our articulations. In order to account for the swiftness with which speech is repaired, it is crucial to find a neurobiological mechanism that can address the delay between initiation and feedback.

Yet this question of how we monitor and control our movements is not isolated to speech. In the production of any action, physical constraints on the speed of neural signal transmission create a delay between the implementation of a motor program and the processing of its sensory consequences (Kawato, 1999). Hickok (2012a) likens this situation to trying to drive a car while only looking through the rear-view mirror; one can only see where the car is on the road and roughly which direction it is pointed in, but not where it is going. Ostensibly, the only way to drive under such conditions is very slowly and constantly checking for changes in the mirror. Yet the speed with which we coordinate our physical movements in order to produce language, whether signed or spoken, suggests that production involves mechanisms that allows us to look forward in addition to looking backward.

Models of motor control suggest that the solution to this problem lies in the ability of the production system to predict the sensory consequences of a given motor program utilizing internal forward models (Davidson and Wolpert, 2005; Jordan and Rumelhart, 1992; Wolpert, 1997). Consider a very simple situation, such as lifting one's hand off of a table. As described over a century ago by William James (James, 1890), action first involves the activation (or "ideation" in James' terms) of a *sensory* target, which then generates a corresponding motor program with the goal of bringing about the desired sensory state. Activation of the motor program generates a forward model of what the physical state of the hand will be after the motor program has been implemented. At the same time, a second corollary forward model (also called an 'efference copy') is activated to predict the sensory consequences of the activated motor program, i.e., what the hand will look like when it is raised (Holst and Mittelstaedt, 1950; Sperry, 1950). The position of the hand is described as its "state". Thus, these 'state feedback control' (SFC) models solve the delay problem by comparing the actual sensory feedback to the predicted feedback. This enables us to rapidly compare the sensory feedback of the hand raising motion to our expectations and correct for deviations from its predicted trajectory (Wolpert and Ghahramani, 2000).

SFC models also provide an account for why delayed auditory feedback can so severely disrupt speech. A delay into the auditory signal creates a discrepancy in the timing of the comparison between efference copy and sensory feedback. SFC models for speech (Houde and Nagarajan, 2011) have been successful in accounting for the suppression of the auditory cortex during speech production (Heinks-Maldonado et al., 2005) as resulting from a match between the efference copy and the perceived signal. Evidence

indicates that this efference copy contains information not only about when an utterance will occur, but how it will sound as well. Experiments with frequency altered feedback have found that participants alter their articulations to compensate for frequency shifted feedback (Houde and Jordan, 1998, 2002; Purcell and Munhall, 2006), and that unexpected shifts in auditory feedback lead to enhanced cortical responses (Chang et al., 2013). More recently, a 'hierarchical state feedback control' model (Hickok, 2012a) has been developed that integrates psycholinguistic models of message formulation (e.g., Dell, 1986; Levelt, 1989; Levelt et al., 1999; Levelt, 1983) with state feedback control models of speech production. This model highlights the importance of sensorimotor processes and hierarchical feedback in order to provide a comprehensive account of speech production.

1.1.2 Sensorimotor experience in speech perception

While there is no doubt as to the importance of sensorimotor experience for the development and maintenance of speech production abilities, it is unclear to what extent such experience may shape speech perception. Early psycholinguistic research noted that children's perceptual or comprehension abilities often outpace their production abilities. In one classic example of this phenomenon, a child points to a fish and says "my [fis]"; when the adult responds by asking "is this your [fis]?", the child shakes their head, but when asked "is this your fish?" the child nods and responds "yes, my [fis]" (Brown and Berko, 1960). This example demonstrates that children are able to perceive phonetic contrasts that they are not yet able to produce and that their substitutions during production do not seem to affect their perception.

Despite this dissociation observed in language acquisition, recent experimental evidence suggests that the production and perception of speech may be tightly linked. For example, a recent study demonstrated that blocking the position of the tongue using a pacifier impedes infants' ability to discriminate non-native phonetic contrasts (Bruderer et al., 2015). Similarly, in adults, stretching a listener's facial muscles during presentation of a speech sound alters behavioral categorization (Ito et al., 2009) and neural responses (Ito et al., 2016) to that sound. In a demonstration of a functional role for sensorimotor experience, Adank et al. (2010) demonstrated that imitating a novel accent increases the intelligibility of sentences spoken in that accent, while passive listening or simple repetition does not. These studies support theoretical stances, such as ideomotor theories, that posit tight links between action and perception (for a more detailed discussion of ideomotor theory, see Ch. 5).

This thesis considers two ways in which sensorimotor experience may affect speech perception. The first is an online mechanism that actively generates predictions during speech perception (Brown and Kuperberg, 2015; Pickering and Garrod, 2007), in the same manner that efference copies predict sensory information during production (Hickok, 2012a; Houde and Nagarajan, 2011). The second is an offline mechanism by which sensorimotor experience reshapes sensory representations or sensory-to-phonetic mappings utilized during speech perception (Schwartz et al., 2012).

1.1.2.1 Sensorimotor predictions shape speech perception

Under difficult listening conditions, foreknowledge about the content of an upcoming spoken word can increase its perceived intelligibility (Sohoglu et al., 2014). This demonstrates that speech perception may be influenced by sensory expectations (Brown and Kuperberg, 2015; Davis and Johnsrude, 2007). In the context of speech production, top-down information is posited to form a forward model about upcoming sensory information. Similar forward models about expected sensory sensations have also been proposed to be active in general perceptual processes (Grush, 2004; Wilson and Knoblich, 2005) and speech perception in particular (e.g. Pickering and Garrod, 2007, 2013; Rauschecker and Scott, 2009).

Pickering and Garrod (2007) argue that listeners covertly imitate or simulate the speaker utilizing their own production systems. Covert simulation utilizes the language production system to generate forward models of the content of upcoming speech at multiple linguistic levels (Pickering and Garrod, 2013). Two notable pieces of evidence lend support to the claim that the production system is specifically implicated in speech emulation. First, infants ability to make early eye-movements to semantically predictable objects on a visual display has been found to correlate with their production abilities, not their comprehension abilities (Mani and Huettig, 2012). Second, visual and audiovisual word recognition has been found to be more accurate when participants are presented with recordings of themselves rather than recordings of other participants (Tye-Murray et al., 2013, 2015). Similar effects in action perception (Knoblich and Flach, 2001; Knoblich et al., 2002) have been put forward as evidence that participants utilize their own ego-centric sensorimotor experience when predicting others actions (Wilson and Knoblich, 2005). When a perceived stimulus is more similar to how the viewer or listener would produce that same stimulus, the match between prediction and sensation is greater and recognition is facilitated.

1.1.2.2 Sensorimotor experience maps acoustic information to linguistic representations

While the previous section examined how sensorimotor processes active during speech production may also influence speech perception via efference copy, this section examines to what extent repeated sensorimotor experience may shape representations of speech sound categories. In the (H)SFC (Hickok, 2012a; Houde and Nagarajan, 2011) model, producing speech first requires an inverse mapping from the desired sensory outcomes to a corresponding motor program. This requires learning the mapping between a distal environmental outcome and a proximal action (Jordan and Rumelhart, 1992). As vocal motor learners, human infants acquire speech sound templates as a result of exposure to their native language (Doupe and Kuhl, 1999). At approximately seven months of age, "canonical babbling" begins, during which random sound production leads infants to learn the sensory-motor transform between an acoustic signal and a corresponding motor program (Guenther and Vladusich, 2012). This active sensorimotor exploration appears to be crucial in order to develop the ability to speak. Songbirds who have had sufficient tutoring experience to acquire a mature sound template nevertheless fail to produce natural song when deafened during this sensorimotor learning phase (Konishi, 1965), and similar effects have also been observed in pre-pubescently deafened human children (Plant and Hammarberg, 1983). Thus, becoming a speech producer critically relies on learning sound-to-motor mappings through sensorimotor experience. While plasticity decreases with ageing, exposure to altered feedback can disrupt sensorimotor mappings in both adult birds (Brainard and Doupe, 2000; Leonardo and Konishi, 1999) as well as adult humans (Houde and Jordan, 1998, 2002), suggesting that sensorimotor feedback continues to be used in adulthood to maintain production abilities (Lane and Webster, 1991; Niziolek et al., 2013; Sitek et al., 2013).

These data imply two possible ways in which sensorimotor experience may affect representations utilized in perception. First, if speech perception involves accessing articulation-based representations to decode speech sounds, as has been proposed by the motor theory of speech perception (Galantucci et al., 2006; Liberman et al., 1967) and analysis-by-synthesis (Poeppel et al., 2008; Stevens and Halle, 1967), then altering the mapping from sound category to articulation to acoustics may also work inversely to change the mapping from acoustics to sound category. Alternatively, repeated action-sensation chains (such as in the case of babbling) generated by sensorimotor experience may lead to coding of representations in terms of their sensory outcomes (Greenwald, 1970; Hommel et al., 2001), or generation of higher order 'common' representations in which there is no distinction between the two modalities (Prinz, 1990). The DIVA model of

speech sound acquisition suggests that common representations may form a "speech-sound map", located in left ventral premotor cortex and left inferior frontal cortex, containing cells that respond both to the production and perception of certain speech sounds (Guenther and Vladusich, 2012; Tourville and Guenther, 2011). Sensorimotor experience may thus alter motoric or shared representations possibly involved in speech perception.

Second, perceptual learning experiments (Samuel and Kraljic, 2009) have shown that speech sound categories can be retuned on the basis of subsequent feedback (e.g., indicating whether a response was "correct"; Lametti et al., 2014a), lexical context (Eisner and McQueen, 2005; Norris et al., 2003), and visual information (Bertelson et al., 2003). Even if perception does not make reference to articulatory representations, sensorimotor experience may act as a sort of 'supervised learning' to retune speech sound categories (Jordan and Rumelhart, 1992). If a speaker's productions consistently vary (e.g., in response to altered feedback; Houde and Jordan, 2002), rather than continuously perceiving the productions as errors, the distribution of the category may shift such that these productions are now "on-target" (Kleinschmidt and Jaeger, 2015a). Thus, sensorimotor experience may retune speech categories later utilized in perception (Lametti et al., 2014b; Shiller et al., 2009).

1.1.3 Methodology

Sensorimotor experience may affect perception A) via efference copies generated from motor areas about the sensory consequences of *predicted* speech, or B) by retuning sound representations or mappings between acoustic and articulatory representations. Teasing apart exactly which of these processes may be occurring is likely to be as difficult as distinguishing between prediction and integration, and is beyond the scope of this thesis. Indeed, recent accounts of analysis by synthesis (Poeppel et al., 2008) suggest that articulatory representations are implemented via efference copies, thus implicating both mechanisms at once.

This thesis utilized two primary methodologies in order to examine the role of sensorimotor experience in speech perception. First, I utilized altered auditory feedback (AAF) in order to alter participants' motor-to-acoustic mappings during production and then examined how this may have altered their behavior in subsequent phonetic categorization tasks. The frequency alteration devices utilized in Chapters 2 and 3 were developed by John Houde (Houde and Jordan, 1998, 2002) and Shanqing Cai (Cai et al., 2008; Tourville et al., 2013) respectively, and both operate under the same functional principles. A participant is fitted with headphones and a microphone. Input from the

microphone is transmitted to a computer sound card for spectral analysis and formant tracking. The digital signal can then be resynthesized in a manner allowing for manipulation of different acoustic parameter, such as formant frequency. The manipulated signal is then fed back to the participant's headphones, with an overall delay of approximately 6-12ms. This allowed us to introduce a discrepancy between expected and actual sensory feedback. In response to such alterations, participants have been found to exhibit compensatory responses opposite the direction of the shift in feedback (Houde and Jordan, 2002; MacDonald et al., 2011). If these shifts are introduced gradually over a number of trials, speakers often fail to notice any discrepancy in their production. After a certain point, participants' compensatory articulations tend to stabilize, at which point participants are said to have "adapted" to the new sensorimotor mapping. This is thought to be a true remapping, as it is found to persist when feedback is masked by noise (Villacorta et al., 2007). When the feedback alteration is removed, articulation slowly returns to normal (though after-effects may persist; Purcell and Munhall, 2006). Recent experiments that have examined the effects of adaptation on perception have found shifts in the phonetic categorization of fricatives (Shiller et al., 2009) as well as vowels (Lametti et al., 2014b). This method is therefore eminently suitable for studying how sensorimotor experience may affect phonetic categorization.

The second methodology utilized in this thesis is open-response word recognition, specifically examining recognition of self-produced stimuli (Chapters 4 and 5) and stimuli produced by a range of talkers (Chapter 5). An open-response format was chosen to maximize the ecological validity of the recognition task (Clopper et al., 2006). Word recognition was specifically chosen in contrast to phonetic categorization because evidence suggests that these two "speech perception" tasks involve different neural processes (Hickok and Poeppel, 2000). For example, utilizing the same stimuli and response options for two tasks, Krieger-Redwood et al. (2013) found that transcranial magnetic stimulation of premotor cortex decreased reaction times during a phonetic categorization task but not during a semantic categorization task. Blumstein (1994) found that participants with aphasia exhibited relatively intact word recognition, yet were greatly impaired in syllable identification. Similarly, Stasenko et al. (2015) recently reported a case study in which a stroke patient exhibited a normal categorical perception boundary when the task involved discriminating whether two words differed acoustically, yet was not able to phonetically label the same stimuli in an identification task. These dissociations have led to dual-stream models of speech perception and production (Hickok and Poeppel, 2004, 2007), with a ventral stream dedicated to lexical processing and a dorsal stream subserving sensorimotor integration. While there is debate as to the relative involvement of each stream during speech perception (Poeppel and Monahan, 2008; Rauschecker and Scott, 2009) as well as the importance of sub-lexical processes to

everyday word recognition (Hickok, 2014), these findings clearly indicate that phonetic categorization differs from other linguistic tasks. Therefore, in order to generalize any experimental findings to the role of sensorimotor experience in everyday interactions, it was crucial to utilize a task that is more likely to tap into processes associated with lexical access.

1.1.4 Thesis Outline

Chapters 2 and 3 examined the role of sensorimotor experience in phonetic categorization, utilizing altered auditory feedback. Chapters 4 and 5 examined the role of sensorimotor experience in word recognition.

Chapter 2 investigated to what extent altering the mapping between articulation and acoustics during production alters subsequent speech perception. In this regard, sensorimotor experience informs the listener as to how a continuous acoustic value maps onto a discrete phonetic category (i.e., categorical perception). While previous experiments suggest that this process does in fact occur (Lametti et al., 2014b; Shiller et al., 2009), in each of these experiments the speech sound produced during the motor adaptation task was the same sound presented during the phonetic categorization task. Thus, this does not rule out the possibility that changes in perception were driven by auditory perceptual learning (Samuel and Kraljic, 2009) or response bias rather than altered sensorimotor mappings. To rule out this possibility, I examined sensorimotor adaptation effects on perceptual compensation for coarticulation (e.g., Mann and Repp, 1980; Repp and Mann, 1978; Whalen, 1981, *inter alia*). Due to coarticulatory effects on production, the acoustic realization of a phone is often dependent on its phonetic environment, such as the centroid frequency of a fricative being dependent on the quality of a following vowel (Kunisaki and Fujisaki, 1977). Perceivers appear to be aware of these effects, and will categorize an ambiguous fricative differently depending on its vowel context. I capitalized on this phenomenon by testing whether the perception of a fricative is affected by sensorimotor adaptation of a following vowel. If sensorimotor adaptation causes a certain acoustic value to be mapped onto a different place of articulation, then this ought to result in changes to perceptual compensation for coarticulation effects following adaptation.

Chapter 3 examined neural correlates of production-perception interactions in the phonetic categorization of vowels. During phonetic categorization, vowels of differing height (i.e., with differing degrees of jaw closure) elicit distinct electrophysiological responses (Obleser et al., 2003b; Shestakova et al., 2004). Bidelman et al. (2013) found

that phonetic categorization of speech sounds correlated with variation in the P2 component of the auditory event-related response (ERP). This same component has recently been found to be sensitive to sensorimotor adaptation (Ito et al., 2016). I investigated how sensorimotor adaptation may alter cortical ERPs during phonetic categorization. In addition, I examined correlations between changes in production, phonetic categorization, and cortical responses to determine to what extent changes in phonetic categorization can be characterized as a remapping between acoustics and articulation, and how this remapping may alter neural responses during perception.

Chapter 4 investigated to what extent sensorimotor experience may be involved in word identification. Motor facilitation during phonetic categorization has been found to be modulated by the perceived similarity between the sound presented as a stimulus and the same sound produced by the listener (Bartoli et al., 2015). In non-speech experiments, participants have been found to exhibit 'self-advantages' when performing perceptual tasks contrasting self-produced stimuli and stimuli produced by other participants (Knoblich and Flach, 2001; Knoblich et al., 2002). Similarly, self-advantages in word identification have been found for visual (Tye-Murray et al., 2013) and audiovisual (Tye-Murray et al., 2015) speech stimuli. This may be evidence that representations utilized during perception are constituted by one's own sensorimotor experience (Hommel et al., 2001; Prinz, 1990), or that emulatory processes recruit one's own motor expertise during perception (Wilson and Knoblich, 2005). In this study, I tested participants' recognition of degraded words produced either by themselves or by one average speaker drawn from the participant sample. This average speaker constituted a fair standard against which to test self-advantages in word recognition. I predicted that if perceptual representations are more reflective of a participant's sensorimotor experience, then word identification should be facilitated for self-produced stimuli.

Chapter 5 built upon the results of Chapter 4 by comparing the influence of sensory experience and sensorimotor experience in determining talker-listener intelligibility. Spoken language consists of individual exemplars of phonetic categories (e.g., vowels, syllables) that comprise statistical distributions over acoustic dimensions such as pitch, formant frequency, duration, etc... These categories thus have a rich internal structure that listeners are sensitive to (Clayards et al., 2008). If representations utilized for speech are reflective of sensory input across a wide range of talkers, then we may expect word identification to be easier when a talker produces speech that more closely approximates the prototypical values of the target categories. Alternatively, if representations are more reflective of one's own sensorimotor experience, then we would expect word identification to be easier when the talker and the listener produce statistically similar speech sounds. I contrasted these alternative hypotheses by presenting participants with sentences produced by seven talkers (one of which was the participant themself), and asked

participants to identify two key words in each sentence. This enabled us to both quantify the intelligibility of each talker to each listener as well as measure self-advantages in word recognition. Phonetic analysis of the recordings then enabled us to test to what extent prototypicality (as a correlate of sensory experience) or similarity (as a correlate of sensorimotor experience) could predict talker intelligibility.

Chapter 6 summarizes and discusses the results of the four experimental chapters. For Chapters 2 and 3, in which participants performed phonetic categorization tasks, I discuss how phonetic categorization may reflect a mapping between acoustics and articulation in light of probabilistic models of sound categorization. For Chapters 4 and 5, I discuss the relative weighting of sensory and sensorimotor experience, how this may be shaped by an individual's history of sensory input, and how the involvement of sensorimotor experience may rely on the presence of different acoustic cues. The results of all chapters are then examined with respect to recent neurobiological models of speech perception and production. Finally, I consider the limitations of the thesis and outline possible avenues for future research.

Chapter 2

Sensorimotoric Adaptation Affects Perceptual Compensation for Coarticulation

A given speech sound will be realized differently depending on the context in which it is produced. Listeners have been found to compensate perceptually for these coarticulatory effects, yet it is unclear to what extent this effect depends on actual production experience. In this study, we investigate whether changes in motor-to-sound mappings induced by adaptation to altered auditory feedback can affect perceptual compensation for coarticulation. Specifically, we tested whether altering how the vowel [i] is produced can affect the categorization of a stimulus continuum between an alveolar and a palatal fricative whose interpretation is dependent on vocalic context. We found that participants could be sorted into three groups based on whether they tended to oppose the direction of the shifted auditory feedback, to follow it, or a mixture of the two, and that these articulatory responses, not the shifted feedback the participants heard, correlated with changes in perception. These results indicate that sensorimotor adaptation to altered feedback can affect the perception of unaltered yet coarticulatorily-dependent speech sounds, suggesting a modulatory role of sensorimotor experience on speech perception.

2.1 Introduction

The drive to find parity between production and perception reflects the fact that humans are both producers and perceivers of speech (Casserly and Pisoni, 2010). One of the central questions in speech perception is whether and to what extent our perception of an acoustic speech signal maps onto the physical mechanisms utilized in order to produce that sound (Liberman and Mattingly, 1989). While some theories claim that the representations accessed during speech perception can be described succinctly in terms of acoustics (e.g. Blumstein and Stevens, 1981), others posit that speech perception involves accessing representations more directly related to the articulatory gestures that generated the speech (e.g. Fowler, 1986; Liberman and Mattingly, 1985; Poeppel and Monahan, 2011). This study sought to contribute to this debate by asking whether altering speakers articulation-to-sound mapping for the production of a vowel has consequences for their perception of coarticulated consonants whose interpretation is dependent on vowel context.

While many phoneticians and psycholinguists describe the representations utilized for perception in articulatory terms, researchers in the field of speech motor control have, somewhat paradoxically, as Hickok et al. (2011) have noted, more consistently characterized speech production not as implementing rigid articulatory programs but as attempting to hit an acoustic or somatosensory target (Houde and Nagarajan, 2011). In order to accurately produce the intended target, the speaker must map between an articulatory program and its expected sensory outcomes. However, such models suggest that the articulatory sequences themselves are flexible and can be changed in order to generate a particular acoustic pattern. The stability of an articulatory motor program therefore rests only on its ability to consistently generate intended sensory targets. Furthermore, if speech perception involves mapping from acoustics to articulation, altering this mapping in a production task should alter speech perception as well.

Substantial evidence for a sensory-centric view of speech production stems from experiments utilizing Altered Auditory Feedback (AAF) devices (Houde and Jordan, 1998, 2002) that enable researchers to independently manipulate spectral and temporal properties of a speaker's voice in real time. In response to repeated and consistent perturbations of auditory feedback, speakers alter their productions to more closely approximate their intended sensory outcomes, and this adaptation effect persists even when altered feedback is replaced by masking noise or removed (Purcell and Munhall, 2006). Sensory targets need not necessarily be acoustic, but may also be somatosensory, as similar experiments using altered somatosensory feedback suggest (Lametti et al., 2012).

In addition to investigating speech motor control, this technique has also probed the relationships between production and perception. For example, participants who show more acute discrimination of first formant differences are also found to compensate more in response to perturbations of that formant (Villacorta et al., 2007). Conversely, other experiments suggest that adaptation to AAF may also alter the way speech is perceived, possibly due to a 'restructuring' of the motor-to-acoustic mappings.

In a study by Shiller et al. (2009), for example, participants were asked to produce [s]-initial CV or CVC words under conditions of altered (AF) or unaltered (UF) auditory feedback. Prior to and following speech training, both groups completed a phoneme identification task consisting of stimuli along a continuum between "a said" and "a shed". Compared to the pre-test, the AF group reported more instances of [s], while the UF group reported fewer instances of [s]. This suggests that the changes in the representations accessed during phoneme categorization had been altered by the participants' experience during the production task. This claim was further supported by the results of a passive listening group which listened to an average participant from the AF group, but showed no difference between pre- and post-exposure phoneme categorization.

Adaptation to AAF can also affect subsequent vowel perception. Lametti et al. (2014b) performed two experiments in which participants were tested on categorization of vowel stimuli between "head" [ɛ] and "had" [æ] (Exp. 1) or "head" and "hid" [ɪ] (Exp. 2). The purpose of the experiment was to determine which of these influences, auditory or articulatory, would lead to a change in the categorization of the test stimuli. Participants were separated into two groups, differing in the direction of the feedback perturbation. All participants were tasked with producing the word "head," containing the vowel [ɛ]. In one group, the frequency shifted the vowel in the direction of [æ], in which case the participants would have to articulate a more [ɪ]-like vowel in order to compensate for the shift; the direction of the perturbation was reversed in the other group. In both experiments, it was found that only the group that *articulated* into the phonetic test continuum region (as a result of adaptation) showed significant changes in vowel categorization. In Exp. 1, only participants who had compensated for the shifted feedback by articulating a more [ɪ]-like vowel during the production of the word "head" demonstrated a change in the perception of a continuum between [ɛ] and [ɪ], while in Exp. 2, it was found that only participants who articulated a more [æ]-like vowel demonstrated a shift in perception. The authors conclude from this that the shift in perception follows the direction of the articulation rather than the acoustic input. As in Shiller et al. (2009), the possibility that this effect was due to simple auditory exposure to shifted feedback was ruled out by inclusion of a passive listening control group.

These two studies (Lametti et al., 2014b; Shiller et al., 2009) suggest that altering motor-to-auditory mappings can alter the perception of speech sounds. In both studies, the authors suggest that their findings lend support to the idea that production and perception are closely linked, and that the motor system plays an active role during perception. However, in both of these studies, the speech segment utilized in the adaptation phase was the target segment in the perception phase. Therefore, it is unclear to what extent such production-induced shifts are simply a result of a bias induced by the altered feedback procedure or represent true perceptual changes induced by auditory-motor remapping.

By using altered auditory feedback to study perceptual compensation for coarticulation – where the segment that is adapted in production can be dissociated from the segment that is tested in perception – it becomes possible to test whether perceptual changes are more than response bias. It has long been observed that the articulation of a speech sound is, as a rule, extremely influenced by its surrounding context, and the same acoustic signal can be widely interpreted based on the context in which it is produced (Liberman et al., 1952). It is the rule, rather than the exception, that segments blend into one another without clear moments of delineation.

> *Nirgendwo ist im Inlaut etwass von Abglitt und Anglitt zu finden, sondern eine wonderbare Koartikulation...* "There is nowhere in its content something of an on-glide or off-glide to be found, only a wonderful coarticulation..." (Menzerath and de Lacerda, 1933, p. 52).

An example of the coarticulated nature of speech can be found in vowel-consonant coarticulation; in English, a vowel preceding a nasalized consonant will also tend to be nasalized (Bell-Berti and Krakow, 1991). Both behavioral (Fowler and Brown, 2000) and neuroimaging experiments (Flagg et al., 2006) reveal that English listeners are sensitive to this nasalization as an indicator of an upcoming nasal consonant. The ability of listeners to recognize segments as the same under such varying conditions, and to utilize such cues, was a primary factor leading to the characterization of representations involved in speech perception as ultimately articulatory in nature (Liberman and Mattingly, 1985).

One particular phenomenon that has been repeatedly investigated with regard to the interactions between coarticulation and perception is known as "compensation for coarticulation" (*CFC*). In addition to being sensitive to acoustic cues for coarticulated segments, listeners have been found to "undo" common coarticulatory effects in a manner that alters their perception of a given segment in a particular context. For example, in the study by Fowler and Brown (2000) on the perception of vowel-nasal consonant

sequences, listeners were found to perceive a nasalized vowel as *less* nasal if it was followed by a nasal consonant compared to when it was followed by an oral consonant. Perceivers perceptually "compensated" for the coarticulatory effects of the nasal consonant on the preceding vowel by attributing acoustic information from one segment to the following segment. Subsequent research has identified a range of contexts in which *CFC* effects occur (e.g. Elman and McClelland, 1988; Fowler, 2006; Lotto and Kluender, 1998; Mann and Soli, 1991; Mann and Repp, 1980; Mitterer and Blomert, 2003; Repp and Mann, 1981), including synthesized, natural, and even non-speech contexts.

Purely perceptual experiments have attempted to disentangle articulatory and acoustic accounts for *CFC* by demonstrating that non-linguistic stimuli may also affect speech categorization (Holt, 2005) or by utilizing stimuli that make different predictions based on articulatory or acoustic information (Viswanathan et al., 2010). In this experiment, however, we attempt to more directly probe the interactions between production and perception by altering the relationship between articulation and acoustics for a particular speech sound and then examining whether this remapping may affect the *CFC* response. If adaptation to AAF involves sensorimotor remapping, then generalization of the remapping to the unadapted speech sound would suggest that the perception of continuous speech involves active utilization of articulatory knowledge to classify speech sounds. Furthermore, testing *CFC* responses on unadapted segments rules out the possibility that any observed effects can be attributed to response bias.

2.2 Experiment 1

A clear *CFC* effect that has been demonstrated in the literature is the effect of vowel quality on the categorization of a preceding fricative (Kunisaki and Fujisaki, 1977; Mann and Repp, 1980; Whalen, 1981). For example, when producing the word "sheep", the tongue position for [i] is already being prepared during the articulation of the word initial fricative. This has the effect of raising or lowering the centroid frequencies of the coarticulated fricative dependent on the quality of the following vowel. The centroid frequency of a fricative will tend to be higher before [i] and lower before [u]. Therefore, in fricative vowel sequences, a certain portion of the "lowness/highness" of the centroid frequency of the preceding fricative can be attributed to the speaker's preparation to articulate the following vowel (in the same manner that nasality on a vowel can be attributed to a following nasal consonant). Early research by Kunisaki and Fujisaki (1977) found different perceptual responses to fricative stimuli along a continuum between [s] and [ʃ] dependent on the quality of the following vowel. These responses were in opposition to the articulatory effects; listeners were more likely to categorize an

ambiguous fricative between [s] and [ʃ] as [s] in the context [-u], and [ʃ] in the context [-i]. This suggests that listeners perceptually compensate for the effects of coarticulation on the acoustic realization of the intended phones.

In the first experiment, we investigate whether changes in the articulation of the vowel [i] due to exposure to AAF may lead to a change in the perception of an unshifted yet contextually dependent fricative continuum between "see" and "she". In two sessions, we asked participants to categorize fricative-vowel stimuli after short periods of production training under conditions of altered (AF) and unaltered (UF) auditory feedback. This within-subjects design enables us to compare the same participants under conditions of unaltered and altered feedback. In AF sessions, participants' auditory feedback in response to their productions of words containing [i] was shifted to more closely approximate their average productions of the vowel [u].

While for certain participants this shift involved some change to the first formant, corresponding to vowel height, the majority of the shifted feedback was confined to altering the second formant, which corresponds to vowel frontness/backness. According to the results of Kunisaki and Fujisaki (1977), as the context-vowel becomes more [u]-like, the proportion of stimuli perceived as [ʃ] should decrease.

To oppose this shift, participants would need to hyper-articulate their productions of the vowel [i], in articulatory terms increasing vowel "fronting" and in acoustic terms increasing the frequency of the second formant. However, it has been found that adaptation to AAF is not total; participants' compensatory articulations only reduced the shift in auditory feedback by approximately 20% (Katseff et al., 2012; MacDonald et al., 2011). Therefore, while participants may oppose the shift induced by AAF, they are still being exposed to stimuli that are shifted compared to an average production. If participants perceptually compensate on the basis of the acoustics they hear during the AF production session, then participants should subsequently report fewer "she" responses in the perception test. Yet in terms of the articulation of the vowel, motoric adaptation leads the vowel to become more [i]-like, in which case we would expect an increase in the amount of stimuli subsequently perceived as "she". As in Viswanathan et al. (2010), the acoustic and articulatory accounts make opposite predictions as to the direction of the perceptual change.

However, this does not rule out the possibility that participants may fail to produce opposing articulatory responses upon exposure to altered feedback. A meta-analysis conducted by MacDonald et al. (2011) of seven AAF experiments (116 participants) found a range of inter-speaker variability in vocal responses to the auditory feedback. In all data analyzed, the target word to be produced was "head", and the shift consisted

of an increase in F1 by 200Hz and a decrease in F2 by 250Hz. Measuring the difference between the last 15 utterances of the baseline phase and the last 15 utterances of the full shift phase, the authors found that 14 of the 116 talkers exhibited a following response in F2, with an additional 4 exhibiting a following response in both F1 and F2 (15% of participants). If all the participants in the present study exhibited such following responses, the acoustic and the articulatory accounts would predict the same result, namely, a decrease in the amount of "she" responses. However, if some participants exhibit following responses and some exhibit opposing responses, an articulatory account would predict a difference between participants exhibiting an opposing response to the AAF compared to participants exhibiting a following response, while the acoustic account predicts that both groups would report more stimuli as "she".

2.2.1 Participants

24 North American native speakers of English (14 = female). Average age at time of testing was 27.42 years (min = 21, max = 36), and all were residents of the Northern California Bay Area at time of testing.

2.2.2 Ethics Declaration

Ethical approval for this study was obtained from the Institutional Review Board of the University of California, San Francisco (Exp. 1), as well as the Ethics Committee of the Social Sciences Faculty of Radboud University (Exp. 2). Written consent was obtained from each participant on the first day of the study. Participants were informed that their participation was voluntary and that they were free to withdraw from the study at any time without any negative repercussions and without needing to specify any reason for withdrawal. All participants were reimbursed for their participation.

2.2.3 Design

The experiment comprised two sessions, an unaltered feedback (UF) and an altered feedback (AF) session. All participants completed both sessions. The UF session always preceded the AF session, to prevent any possible carryover effects from the altered feedback. Each session was separated by a minimum of two weeks.

Within each session, participants performed two tasks, a speaking task and an identification task, separated into blocks. In the speaking task, participants read aloud simple

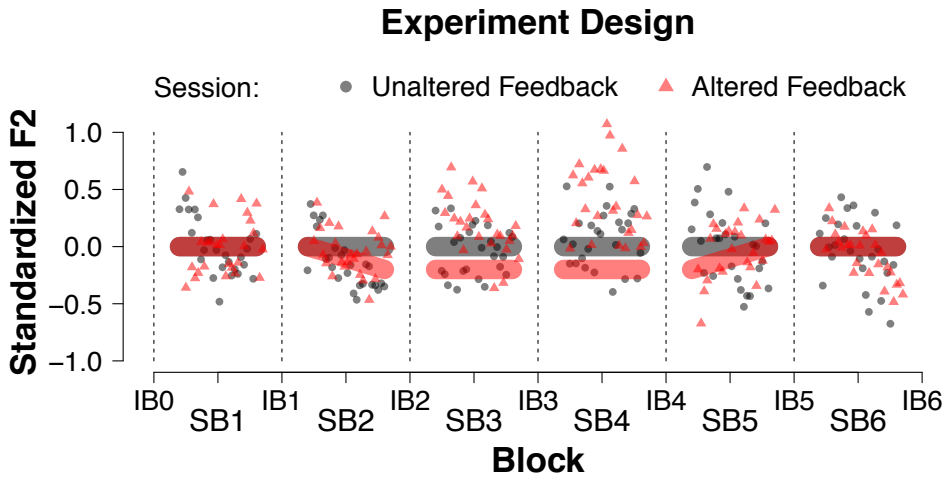


FIGURE 2.1: **Overview of Experiment Design.** Each session consisted of alternating Identification Blocks (IB), and a Speaking Blocks (SB). Grey and red bars indicate auditory feedback in the unaltered (UF) and altered (AF) sessions, respectively. In the AF session, auditory feedback was shifted 20% of the participant's acoustic distance towards the participant's average [u] values, with ramp-up, hold, and ramp-down phases. Grey circles (UF session) and red triangles (AF session) denote grand average standardized F2 by trial.

CVC words. In the perception task, participants listened to synthesized stimuli ranging along a continuum between clear "see" and clear "she", and reported via button press which word they thought they had heard. Both sessions consisted of seven Identification Blocks (IB) and six Speaking Blocks (SB), presented in alternation (Fig. 2.1). A session began with an IB, followed by a SB, and continued in this manner until all seven IBs and SBs had been completed. All participants completed both sessions.

2.2.4 Speaking Task

2.2.4.1 Equipment and AAF signal processing

For both sessions, participants were seated in front of a laptop, and fitted with Beyerdynamic DT 770 PRO noise-isolating headphones. Speech was recorded utilizing a Micromic.C 520 VOCAL head-mounted microphone. Microphone input was routed through an RDL HR-mP2 Dual Microphone Preamplifier to an M-Audio Delta 1010 external sound card.

In order to determine the most accurate audio processing settings to apply altered feedback, each participant was first recorded producing the words "heed", "who'd", and

"had", corresponding to the point vowels [i], [ɛ], and [æ], respectively. Spectral measurements using varying levels of LPC coefficients and frequency cutoff levels were then generated, and the number of coefficients and the frequency cutoff level giving the best formant tracking were selected by visual inspection of the results. Following calibration, participants were recorded producing the same words three times each. The averages of the first and second resonant frequencies for the point vowels [i] and [u] were then utilized as the basis for the frequency shifted feedback (see Procedure).

The input signal from the microphone was analyzed by a frequency alteration device (FAD) as described by Katseff et al. (2012), based on the method of sinusoidal synthesis (Quatieri and McAulay, 1986). The input signal was recorded with a 32-bit sampling depth at a rate of 11025Hz, generating a frame size of 3ms. This frame was then ported into a 400 sample, 36ms buffer for spectral analysis. The acoustic envelope was converted into a narrow-band magnitude frequency spectrum in order to obtain the spectral envelope from the signal. This spectral envelope was utilized to estimate, and modify, the fundamental and resonant frequencies present in the recorded input. The new narrow band magnitude frequency spectrum was then used as the basis for sinusoidal synthesis, in which each harmonic of the spectrum is represented as a sinusoid. The acoustic signal was then generated by sinusoidal addition. In order to maintain continuity between frames, the preceding 3ms frame was also used as input to the estimation of the current 3ms frame, totaling a 6ms analysis window. Additional processing delays between microphone and headphones led to an overall delay of approximately 12ms, regardless of whether or not feedback had been altered.

2.2.4.2 Speaking Task Stimuli

Stimuli for the speaking task consisted of 13 monosyllabic English words in orthographic form ('peep', 'beep', 'deep', 'keep', 'peat', 'beet', 'bead', 'deed', 'keyed', 'peak', 'beak', 'teak', 'geek'). Words were presented in white 30-point font against a black background. All words began with a voiced or voiceless stop consonant, followed by the vowel [i], and ended with a voiced or voiceless stop consonant. Crucially, no words contained a fricative consonant.

2.2.4.3 Procedure

In a Speaking Block (SB), participants were instructed to read aloud the words presented on screen. The thirteen stimulus words were presented in random order. Each word was

presented twice, with all words being presented at least once before repetition began, totaling 26 presentations per block.

In the UF session, there was no altered feedback except for the downsampling of the signal and the processing delay of 12ms. In the AF session, there was no altered feedback in the first (baseline) SB. Starting in the second SB ("on-ramp"), the participants' auditory feedback was shifted towards their average [u] production. In their first experiment, Purcell & Munhall (2006) applied a fixed-frequency shift of ± 200Hz to each participant's F1, which they argue may have led to differing compensatory responses given the vocal tract parameters of the individual participants. Therefore, following Niziolek and Guenther (2013), we defined custom frequency shifts for each participant, based on per-participant differences in average F1 and F2 for the vowels [i] and [u]. For each participant, maximum feedback was 20% of the distance in F1-F2 space from [i] to [u]. Feedback alteration began at 0% perturbation at trial 1 of SB 2, and reached a maximum of 20% perturbation of both F1 and F2 by trial 26 ("on-ramp"). This averaged -263.29Hz for F2, and 6.67Hz for F1, indicating that the primary dimension of the shift was along F2. Perturbation remained at 20% over blocks three and four. In block five ("off-ramp"), perturbation decreased from 20% at trial 1 to 0% at trial 26. Feedback was unaltered in block six.

2.2.4.4 Acoustic Analysis

Semi-automated formant measurements were conducted with the same DSP-processing software utilized for the feedback alteration, and Matlab (Mathworks, 2012). An algorithm measuring the periodicity of the acoustic signal was first utilized to identify the start and end time of the vowel segment of each recording. Formant tracking results were visually inspected to determine that measurements were taken from vowel midpoint or the closest suitable point.

While the altered feedback was defined in terms of change in F1 and F2, the primary direction of change occurred along the F2 dimension. We therefore excluded F1 measurements from analysis. The acoustic measurements for F2 were converted from Hertz to Mel, a logarithmic frequency scale that more closely approximates human hearing, using the formula $2595 * (log(1 + (F2\_Hz/700)))$.

In order to compare production responses across participants, Mel frequency measurements of F2 were standardized according to the following procedure: For each participant, the first SB of each experimental session was designated as the "baseline" block for that session. The Mel value of each trial was subtracted from the average Mel value of

the baseline, and then standardized by dividing by the standard deviation of the baseline block.

$$F2_{standardized} = (F2 - mean(F2_{baseline}))/sd(F2)_{baseline} \quad (2.1)$$

Therefore, these standardized values represented changes from baseline production values in each session with respect to baseline formant variability. All statistical tests were performed utilizing these standardized measurements.

For comparison with other studies, we also calculated an alternative compensation index according to the following formula:

$$F2_{compensation} = 1 - \frac{(F2 - F2_{baseline}) - shiftF2}{shiftF2} \quad (2.2)$$

Where for a given trial, compensation was defined as the percentage return towards baseline from the shifted formant value. Subtracting the resulting proportion from 1 separates responses opposing the direction of the shift (positive values) from those following it (negative values).

2.2.5 Identification Task

2.2.5.1 Stimuli

Stimuli for the identification task were created according to the following procedure: a female native speaker of North American English producing the sentence "say the word 'see'" three times. After selecting the most natural sounding version, the word "see" was extracted. The duration values for each phone segment and the prosodic contour of the word was extracted utilizing Praat (Boersma and Weenink, 2013) and converted into a text-based format readable by Mbrola (Dutoit et al., 1996), a text-based diphone synthesizer. This method enabled the creation of two endpoint stimuli with identical phone durations and prosodic contours. Sample-by-sample interpolation (McQueen, 1991) was then utilized to create a 100-step continuum between unambiguous "see" and unambiguous "she".

2.2.5.2 Pre-test

Before the first session, participants completed a pre-test, utilizing the same materials as the test phase, in order to determine the stimulus-step at which participants switched from reported hearing "see" to "she", by means of an adaptive staircase procedure. In the odd-numbered trials, a random "filler" endpoint stimulus (0 or 100) was presented, as

during pilot testing it was found that presentation of consecutive ambiguous stimuli led to shifts in the categorical perception boundary. These filler trials did not count towards the staircase procedure results. Even numbered trials first began with presentation of the endpoint stimuli (0 or 100, corresponding to unambiguous "see" and unambiguous "she"). Initial step-size was set at 100; after trial four, step size decreased by half after each reversal until either 12 reversals had occurred or step-size remained at one for three consecutive trials.

2.2.5.3 Procedure

In an Identification Block (IB), participants were instructed to listen to an auditorily presented stimulus and indicate via keyboard whether the stimulus sounded more like "see" (button 1) or "she" (button 2). Participants were instructed to wait until the sound file had finished before responding, and to respond with whichever hand felt most comfortable. As it was essential to the experiment that participants listen to both the fricative and the following vowel in order to assess the *CFC* effect, responses made prior to the end of stimulus presentation were not accepted. In such cases participants were verbally reminded to wait until stimulus presentation had finished before responding. While this eliminated the possibility of comparing reaction times, this method ensured that participants listened to both the fricative and the vowel before responding.

Presented stimuli consisted of $\pm 1, 3, 5, 7, 9, 11, 13, 15, 17$ steps above and below each participant's pre-test boundary as well as the endpoint stimuli (step 0 and step 100), for a total of 20 stimuli. Stimulus presentation was pseudo-randomized into sets of four, consisting of one stimulus less than 10 steps above pre-test boundary, one stimulus less than 10 steps below, one stimulus greater than 10 steps above, and one stimulus greater than 10 steps below. In an IB, each stimulus was presented twice for a total of 40 presentations per IB, with all stimuli presented once before repetition. Responses in the Identification Task were coded as 0 (for "see") and 1 (for "she").

2.2.6 Data exclusions

Due to software failure, two participants were unable to complete IBs five and six and SB six of the UF session. When possible, all remaining data from these participants were included in the analyses.

Out of 7436 total trials, 297 trials (4% of total data) were excluded prior to acoustic analysis due to either recording error (e.g. wrong word spoken or spoken outside of

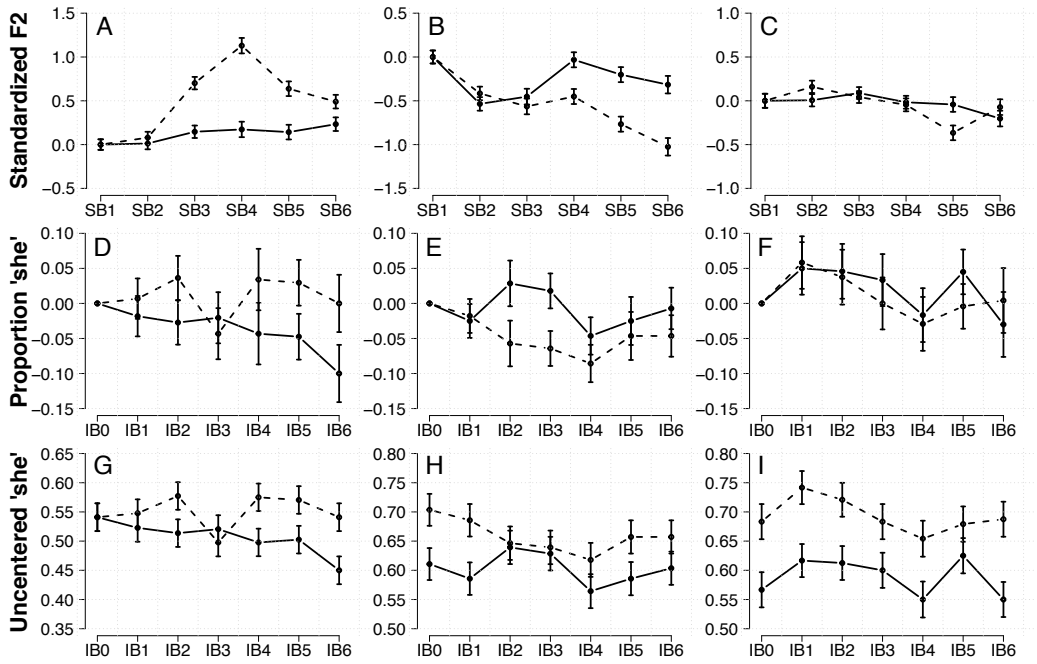


FIGURE 2.2: **Average standardized F2 and proportion 'she' responses by type of participant.** Solid lines refer to the UF session, dashed lines refer to the AF session. Panels A-C represent average standardized F2 for opposers(A), followers (B), and mixed (C) participants. Panels D-F represent group-averages of baseline-centered proportions of 'she' responses for opposers(C), followers (D), and mixed (F). Panels G-I represent group-averages of uncentered proportions of 'she' responses for opposers (G), followers (H), and mixed (I).

recording window) or failure of the formant tracker to locate a stable point for formant measurement. In the remaining trials, each participant's formant measurements were visually inspected for formant tracking errors. Five trials were removed in which formant values for F1 were exceedingly large (greater than 5 standard deviations from centered mean of all participants). We then excluded trials in which F2 was greater than 3 standard deviations from the mean of all participants of that gender (15 trials). Finally, values for F1 and F2 were centered and scaled on a by-participant basis, and based on visual inspection of the data, a conservative value of $\pm 4sd$ was set as the cutoff at which extreme values would more likely have arisen due to tracker error. In total, 41 trials were excluded from analysis in this manner, comprising 0.5% of the data.

2.2.7 Results

2.2.7.1 Production Results

Inspection of individual results in the AF session revealed that, while many participants appeared to oppose the AAF, a number of participants exhibited a decrease in F2 compared to baseline. This suggests that in contrast to an opposing response, certain participants responded to the altered feedback by producing a "following" response in the same direction as the shifted feedback (MacDonald et al., 2011).

We therefore conducted a simple post-hoc division of participants into three groups based on whether standardized F2 averaged over the third and fourth blocks of the AF session, during which the altered auditory feedback was held constant at its maximum value, was greater than baseline for both blocks, less than baseline for both blocks, or above baseline in one block but below baseline in another. Standardized F2 was above baseline for 11 participants ($mean\ of\ blocks\ 3\ and\ 4\ = 0.917, sd = 1.353$) and below baseline for 7 participants ($mean\ of\ blocks\ 3\ and\ 4\ = -0.508, sd = 1.128$), and mixed for 6 participants ($mean\ of\ blocks\ 3\ and\ 4\ = -0.003, sd = 0.867$). We classified these participants as "opposers", "followers", and "mixed", respectively. Average standardized F2 for the three groups in the UF and AF sessions are displayed in Fig. 2.2 (panels A-C).

Standardized F2 values were analyzed with linear mixed effects regression in R (R Development Core Team, 2013) using the lme4 package (Bates et al., 2014). This technique is robust to missing data, and allowed us to include data from the two participants who failed to complete the last few blocks of the UF session. Due to the standardization procedure, block 1 contained essentially no variance and therefore was excluded from the analyses. Model fitting began with a maximal model containing fixed effects for "Type" (opposer, follower, mixed), session (AF or UF), and block (1 - 6; as categorical variables), as well as all two-way and three-way interaction terms. We utilized a maximum random effects structure (Barr et al., 2013) including a random intercept for participant as well as random slopes for block and session. Significance of predictors was assessed by conducting likelihood ratio tests between nested models with and without the candidate fixed effect term. Removal of the 3-way interaction term for type, block and session was found to significantly decrease model fit ($\chi^2(8) = 63.116, p < 0.001$). To simplify model interpretation, we subsetted the data and fit separate models for the AF and UF sessions.

For each by-session model, we began with a maximal random effect structure including random intercepts for participant and a random slope for block, and a maximal

fixed effect structure including main effects for type and block as well as the interaction term. Removal of this interaction term significantly reduced model fit in the AF session ($\chi^2(8) = 19.377, p < 0.013$). Model estimates (with p-values obtained by Satterthwaite approximation) are reported in contrast to block 2 of the AF session for the opposer group. For this group, standardized F2 was found to increase significantly in blocks 3 ($Est. = 0.610, SE = 0.158, p < 0.001$) and blocks 4 ($Est. = 1.038, SE = 0.324, p = 0.005$). In block 2, the opposers differ significantly from followers ($Est. = -0.516, SE = 0.236, p < 0.05$) but not from the mixed group. This difference between opposers and followers from block 2 onwards exhibits a significant decrease in block 3 ($Est. = -0.761, SE = 0.253, p < 0.007$) and a marginally significant decrease in block 4 ($Est. = -1.083, SE = 0.520, p = 0.05$). The difference between opposers and mixed becomes significant in block 3 and remains significant until block 6 (all *ps* < 0.05).

In contrast, in the UF session, removing the interaction term ($\chi^2(8) = 5.004, p = 0.75$), the main effects of type ($\chi^2(2) = 2.596, p = 0.27$), and the main effect of block ($\chi^2(4) = 2.756, p = 0.6$) all failed to significantly impact model fit. These combined results indicate that the opposer, follower, and mixed groups do not reliably differ in standardized F2 when feedback is unaltered, but produce differing response patterns when exposed to altered auditory feedback.

As has been found in previous experiments, adaptation was not sufficient to completely counteract the altered feedback (Houde and Jordan, 1998; Katseff et al., 2012; Purcell and Munhall, 2006). In contrast to the average F2 compensation of 23.2% reported for perturbations of [ɛ] (MacDonald et al., 2011), in this experiment opposers altered their productions by an average of only 5.6% in the third and fourth blocks of the AF session (min = 0.5%, max = 18.1%). Followers were found to alter their F2 by only -2.9% in the below-baseline group (min = -5.8%, max = -0.3%).

2.2.7.2 Identification Results

Our key question concerned whether participants' perceptual responses differed with respect to conditions of altered and unaltered production feedback. However, the production results indicate that participants' vocal motor behavior in response to AAF differed, justifying their division into three separate groups. Figure 2.2 (D-F) displays the baseline-centered proportion of "she" responses in each block and session for the opposers, followers, and mixed groups, with uncentered proportions are displayed in panels G-I.

These graphs indicate that, despite the staircase procedure, the proportion of "she" responses in the first block appears to have shifted between the UF and AF sessions for participants in the follower and mixed groups (*mean difference AF-UF:* $opposers = 0.000, followers = 0.093, mixed = 0.117$). Rather than averaging responses in each block over stimuli, binary coded identification responses were analyzed using mixed-effects logistic regression (Breslow and Clayton, 1993; Jaeger, 2008).

The graphs in Figure 2.2 (D-F) suggest that with regards to overall proportion of responses, participants in all three groups tended to perceive more stimuli as "she" in the AF session compared to the UF session. However, because this increase was present from the first block in the follower and mixed groups, this overall increase in "she" responses cannot be attributed to the AAF. To assess the effect of altered auditory feedback on identification, we conducted separate regression analyses examining the effects of block and session within each group. As with the production data, model fitting proceeded by comparison of backwards-fitted nested models.

For opposers, the initial model contained a random effect structure containing random intercepts for participant and item, as well as random slopes for block and session over participant and session over stimulus. Removing the interaction between session and block significantly decreased model fit ($\chi^2(6) = 29.372, p < 0.001$). Parameter estimates (treatment coded with the baseline block (block 0) of the UF session as reference) indicate that in the UF session, the proportion of "she" responses was significantly less than baseline in the sixth block ($Est. = -1.11, Z = -3.37, p < 0.001$. The estimates of the interaction indicate that in the AF session participants tended to perceive more stimuli as "she" in the second ($Est. = 0.717, Z = 2.204, p < 0.028$), fourth ($Est. = 0.856, Z = 2.641, p < 0.009$), fifth ($Est. = 0.845, Z = 2.564, p < 0.011$), and sixth blocks ($Est. = 1.097, Z = 3.322, p < 0.001$) compared to the corresponding blocks in the UF session (with respect to their change from baseline; Fig. 2.2D).

The model for the follower group included fixed effects of block, session, and the interaction term, as well as random intercepts for participant and stimulus with random slopes for session. Removing the interaction between session and block also significantly decreased model fit ($\chi^2(6) = 15.941, p < 0.02$). Inspection of model estimates indicates that in the UF session, the probability of perceiving stimuli as "she" in block 4 was significantly lower than in the baseline block ($Est. = -0.6838, Z = -2.132, p < 0.04$). In contrast to the opposers, model estimates indicated that participants were already more likely to report stimuli as being "she" in the baseline of the AF session than in the baseline of the UF session ($Est. = 1.2496, Z = 2.747, p < 0.007$). This increased tendency to report stimuli as "she" decreased significantly in the second ($Est. = -1.2236, Z =$

$-2.717, p < 0.007$) and third ($Est. = -1.1650, Z = -2.587, p < 0.01$) blocks of the AF session.

For the mixed group, a model containing the same fixed effect structure was fit with random intercepts for subject and stimulus with random slopes for session and block over participant and session over stimulus. As in the production results, removal of the interaction and main effects did not significantly impact model fit (all *ps* > 0.05).

2.2.7.3 Interactions between Production and Identification

As in other experiments (Lametti et al., 2014b; Shiller et al., 2009), no within-session significant correlations were found between production (standardized F2) and perception (centered response). The results from the production and the perception models seem to suggest that the effect of the altered feedback on perception only emerges when considering the differences in production and identification across the two sessions. The within-subjects design enabled a comparison of each participant's behavior under conditions of both altered and unaltered feedback. We computed differences scores for both the production and the identification data: For the production results, we utilized the standardized F2 scores; for the identification results, we baseline-centered each participant's results to the average of block 0 of each session. This allowed us to compare both the identification and the production results with respect to changes from baseline in a given session. Each participant's average standardized F2 for production and average baseline-centered proportion of "she" responses for identification in the AF session were subtracted from the corresponding results in the UF session. These difference scores were found to have a positive correlation (Fig. 2.3; $r(22) = 0.44, p < 0.031$), indicating that a higher standardized F2 value in the AF session compared to the UF session (corresponding to a more fronted [i] vowel) correlates with an increased likelihood of "she" responses in the AF session compared to the UF session. The fact that this correlation is found with regard to by-session and by-participant baseline-centered proportions of "she" removes the possibility that this pattern is simply due to an overall difference in the number of "she" responses in one session compared to another. For blocks three and four, the two blocks corresponding to maximum perturbation in the AF session, differences scores were found to correlate only for the fourth block ($r(22) = 0.47, p = 0.016$), and not for the third ($r(22) = 0.32, p = 0.13$). This concords with what is seen in the group results, in which a sharp drop is seen in the proportion of "she" responses in the opposer group (Fig. 2.2D), even though standardized F2 is above baseline in this block (Fig. 2.2A).

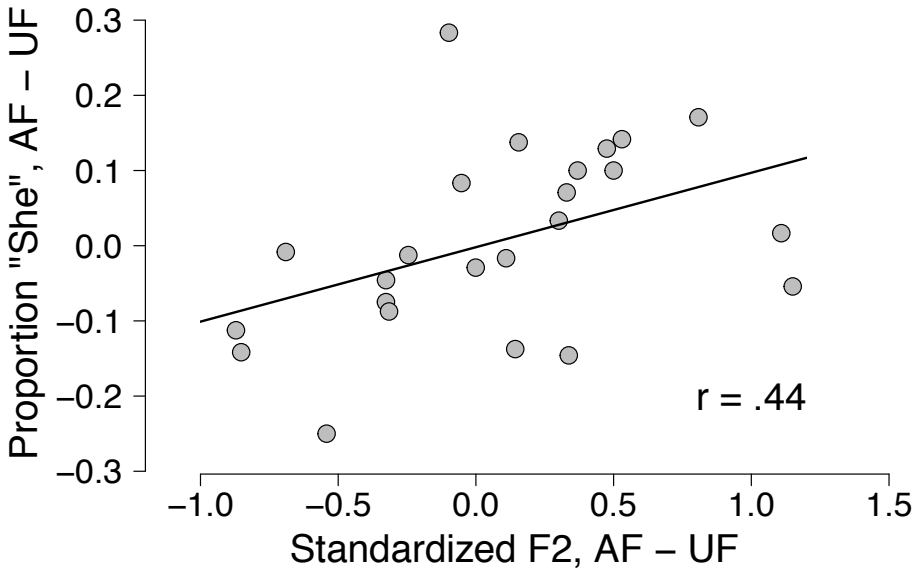


FIGURE 2.3: **Correlation of difference scores.** X-axis denotes, for each participant, the average standardized F2 in the UF session subtracted from average standardized F2 in the AF session. Y-axis denotes average baseline-centered proportion of 'she' responses in the UF session subtracted from those in the AF sesssion.

Inspection of individual by-trial results revealed that some participants exhibited a strong opposing or following response in the initial half of a block, but then returned to baseline in the second half (or vice versa). If vocal behavior in the second half of an SB differed from vocal behavior in the first, this may have "cancelled out" the effect of adaptation on previous trials. If this is the case, then we surmised that stronger correlations between the difference scores may be found if standardized F2 measurements were limited to the second half of each SB (trial > 13). Limiting production data to the last half of the block was found to slightly increase the correlation coefficient ($r(22) = 0.47, p < 0.019$). In addition, we found that the previously non-significant correlation in block three became marginally significant if we limited our measurements to the last 10 trials of the block ($r(22) = 0.41, p = 0.05$). These within-block changes in F2 may explain the sudden drop in identification scores in the third IB despite an increase in average standardized F2 in the corresponding SB.

2.2.8 Discussion

The combined results from the Speaking and Identification tasks suggest that a change in produced F2 correlated with a change in the proportion of stimuli perceived as "she".

Even though participants in the oppose group counteracted only 5.6% of the shifted feedback, they nevertheless exhibited an increase in the proportion of stimuli reported as "she". The direction of the correlation accords with results from similar *CFC* experiments in which a more fronted vowel is associated with more [ʃ] responses (Kunisaki and Fujisaki, 1977; Mann and Repp, 1980). This provides evidence against a purely acoustic account of the results, because each participant (both opposers and followers) heard themselves producing the vowel [i] with a much lower F2 than normal. Instead, the shift in perception correlated with changes in the motoric behavior, supporting an articulatory basis for the perceptual shift. In line with the results of Lametti et al. (2014b), changes in perception were variable and depended on the direction of the articulatory change, rather than universally according with the direction of the shifted auditory feedback.

The relatively large amount of following responses found in this experiment compared to others may possibly be due to the articulatory location of the target vowel ([i]). Acoustically, this vowel occupies an endpoint in terms of both F1 and F2, while articulatorily, the degree of lingual contact is greater for [i] than for non-closed vowels, such as [ɛ], which may increase the importance of somatosensory feedback relative to acoustic feedback (Mitsuya et al., 2015). As in the meta-analysis conducted by MacDonald et al. (2011), we found no significant correlation between a participant's average standardized F2 in the maximum shift blocks and standard deviation for F2 in baseline blocks in either mean centered F2 ($r(22) = 0.031, p = 0.88$) or standardized F2 ($r(22) = -0.036, p = 0.8654$).

2.3 Experiment 2

The results of Experiment 1 suggest that sensorimotoric adaptation in response to altered auditory feedback can affect the perception of an unadapted, yet contextually dependent, fricative continuum. Furthermore, the direction of the observed effects suggests that the change in perceptual function corresponds to the motoric changes rather than the auditory feedback. Previous experiments have found that passive exposure to the recorded speech of a participant compensating in response to AAF fails to change perceptual function, regardless of whether the recorded speech is made by an average compensating speaker (Shiller et al., 2009) or consists of a random selection of stimuli take from several compensating speakers (Lametti et al., 2014b). However, these experiments differed from the present study in that they did not examine how passive listening to such stimuli may affect *CFC*. Therefore in order to determine whether passive listening may also affect phonetic categorization in CFC, we recruited an additional 20 participants to perform a passive listening version of the same task, in which there was

no speech motor activity involved. If we observe no perceptual changes similar to those observed in Experiment 1, this likely indicates that sensorimotor remapping requires experiencing error between an intended sensory target and the articulatory movement enacted to produce that target (Houde and Nagarajan, 2011).

Based on the behavioral division of participants into opposers and followers in Experiment 1, we decided to directly test whether two different types of auditory stimuli may elicit differential perceptual changes. Half the participants were exposed to the recordings of one male participant from Experiment 1, who exhibited an average increase in standardized F2 in response to hearing his own shifted feedback. The recordings consisted of this participant's unshifted input to the AAF device. If the passive listeners in this "unaltered" group perform like the opposers in Exp. 1, we would expect an increase in the number of "she" responses.

The remaining participants were exposed to the shifted auditory feedback that this participant heard during the first experiment (in which the [i] vowel was altered to sound more like [u]). According to an auditory account (Kunisaki and Fujisaki, 1977; Mann and Repp, 1980), the passive listeners in this "altered" group should, if they exhibit a change in perception, report fewer stimuli as "she" compared to baseline.

2.3.1 Participants

Twenty-two North American native speakers of English (all residing in the Netherlands at the time of testing) took part in Experiment 2. Of these participants, two were excluded due to errors in the pre-test boundary finding procedure, in which they responded "see" to almost all stimuli during the test phase. This left twenty participants (M=9, average age = 27.55). Three of the twenty received eight euros for completing the task, while the remainder declined payment.

2.3.2 Materials

Materials for the Identification Task were identical to those used in Experiment 1. Materials for the Passive Listening Task consisted of either the recorded input (32-bit, 11050Hz sampling rate) from one male speaker from Experiment 1 during the AF session (unaltered-group), or the altered auditory feedback which this speaker heard during the experiment (altered-group). As in Shiller et al. (2009), this speaker was selected from amongst the above-baseline participants (the opposers) for exhibiting an average but not extreme increase in standardized F2 in response to the AAF.

2.3.3 Procedure

Participants were seated in a sound-proof booth in front of a computer monitor. The Identification Boundary Pre-Test and Identification Task were identical to that reported in Experiment 1, except that the experimental software used to deliver the stimuli was Presentation instead of Matlab. In contrast to the previous experiment, the Speaking Task was replaced with a Passive Listening Task (PLT), in which participants were instructed to silently read the words that appeared on the screen and listen to the voice as it read each word aloud.

2.3.4 Results and Discussion

As in Experiment 1, the average proportion of [ʃ] responses were centered to each participant's average in the baseline block (block 0). Group averages were calculated from centered averages. Average proportion of "she" responses in the input group was 0.017 (sd = 0.085), while average proportion of "she" responses in the output group was -0.002 (sd = 0.104). As in Experiment 1, results were analyzed using mixed-effects logistic regression. Models with random slopes for block failed to converge, therefore the model contained random intercepts for participant and centered stimulus. Backwards-fitting began with a maximal fixed effects structure including main effects for block and group (input/output) as well as the interaction term. Model comparison revealed no significant interactions, and neither the effect of group ($AIC = 2945.5$, $BIC = 2972$, $LogLik = -1468.8$, $\chi^2(1) = 0.55, p = 0.46$) nor block ($AIC = 2945.9$, $BIC = 3005.6, LogLik = -1464, \chi^2(6) = 10.15, p = 0.11$) was found to improve the null model.

While this accords with previous results (Shiller et al., 2009), the lack of any change in the perceptual boundary may seem rather counterintuitive given research demonstrating that exposure to ambiguous auditory stimuli in the context of biasing lexical information can drive perceptual retuning (Samuel and Kraljic, 2009). In such experiments, participants are exposed to an ambiguous sound that falls between two phonemic categories, such as [f] and [s], in a lexical context that leads the listener to categorize the ambiguous sound as belonging to one of the two categories (Norris et al., 2003). This lexical bias has been found to have a strong effect on how an ambiguous phone is perceived. The resulting effect is that if the lexical context biases listeners to categorize the ambiguous sound as [f], they are subsequently more likely to classify stimuli along an [f]-[s] continuum as [f], while the inverse is found if participants are led to categorize the ambiguous sound as [s]. Further experiments with this paradigm have found that this effect can generalize to unexposed words (McQueen et al., 2006) and

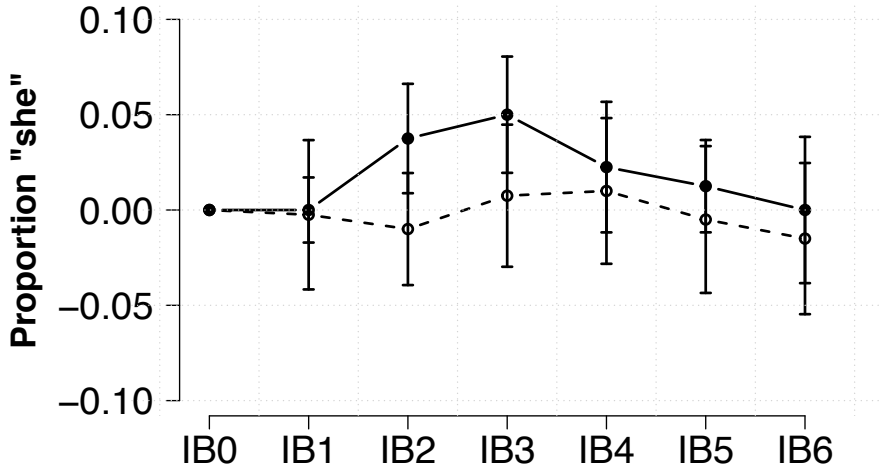


FIGURE 2.4: **Passive listening results.** Filled circles and solid lines indicate average 'she' responses for participants exposed to the unshifted recordings of a model participant from Exp. 1. This model participant exhibited an increase in standardized F2 in response to the altered feedback that was typical of the opposer group. Empty circles and dashed lines denote participants exposed to the shifted feedback that the model participant heard during the AF session.

can remain stable for as long as 12 hours (Eisner and McQueen, 2006), comparable to the durable effects seen in sensorimotor adaptation (Nourouzpour et al., 2015; Ostry et al., 2010). The apparent contradiction between the lexically-guided retuning results and the results of Exp. 2 may be due to the specificity of the adaptation; Kraljic and Samuel (2005) found that perceptually guided lexical retuning for fricatives along an [s] - [ʃ] continuum generalized from a female training voice to a male test voice, but not in the opposite direction. The authors attribute this asymmetry to the fact that the female training stimuli were close to the frequency of the male test items, while the male training stimuli were far from the female test stimuli, suggesting that generalization may depend on acoustic similarity. This provides a reasonable explanation for the fact that we see no effects in Exp. 2, as the silent listening task utilized a male voice while the identification stimuli were based on a female voice.

2.4 General Discussion

The results of the first experiment indicate that participants changed their vocal behavior, as reflected in acoustic measurements of their baseline standardized second formant

(F2), in response to the altered auditory feedback. Our results differ from previous experiments in that, while 11 participants adapted to the feedback by opposing it (i.e., their standardized F2 increased), a relatively large number (13 of 24 participants) failed to oppose the shifted feedback. Instead, in 7 of these participants, standardized F2 was found to decrease relative to baseline, following the direction of the shifted feedback.

Critically, the 11 opposing participants differed from the 7 following participants with regard to their behavior in the identification task. The opposing group reported more instances of "she" in the altered feedback blocks while the following group exhibited a decrease in the amount of stimuli identified as "she". Differences in the speaking task between altered and unaltered feedback sessions were found to correlate with corresponding differences in the identification task, suggesting that responses in the identification blocks were influenced by vocal behavior in the immediately preceding speaking blocks. This suggests that the perceptual processes involved in compensation for coarticulation can be modulated by the observers own sensorimotor experience.

While exposure to nonlinguistic acoustic stimuli is known to affect *CFC* (Holt, 2005), the pattern of results observed in this experiment conflict with a purely auditory explanation. In both groups (opposers and followers), the effect of the altered feedback was to decrease the F2 that participants heard by a substantial amount. As adaptation in articulation only counteracted a small portion of the shift, all participants heard themselves producing the vowel [i] with a lower than average F2. We predicted that if perception is influenced by articulatory behavior rather than acoustic feedback (Lametti et al., 2014b), then the proportion of stimuli perceived as "she" should increase if participants adapted to the altered auditory feedback by increasing the amount to which the intended vowel is fronted (and vice versa). However, if the *CFC* effect is instead modulated by auditory experience, we predicted that in both cases we should expect to observe a decrease in the proportion of "she" responses due to the decrease in heard F2. The articulatory account appears to have been borne out in the results, as exhibited by the different shifts in identification responses between participants who opposed vs. followed the direction of the altered auditory feedback.

Furthermore, shifts in perception were found to correlate with motoric adaptation, not auditory exposure, as found by Nasir and Ostry (2009) and Mattar et al. (2011), and no consistent changes in perceptual function were found in two groups of passive listeners (Exp. 2). However, it should be noted that the correlation between articulatory and perceptual responses was only found when taking into account differences in behavior between unaltered and altered sessions. Similar studies have failed to find correlations between adaptation and perceptual change (Lametti et al., 2014b), leading some to posit

that motoric and sensory adaptation proceed somewhat independently (Nourouzpour et al., 2015).

The pattern of results observed in Experiment 1 seem to accord with gesture-based accounts of compensation for coarticulation (Viswanathan et al., 2009, 2010). As has been suggested in previous research on adaptation to altered feedback in production (Jones and Munhall, 2005), we propose that adaptation results in a remapping between an articulatory/somatosensory representation of a phonetic target and its acoustic consequences. The results of the experiment suggest that during the identification task, the listener's remapping between the articulation or phonetic category onto the acoustics of the context vowel results in a different phonetic categorization of the same fricative centroid.

Prior to exposure to altered auditory feedback, participants have a relatively stable mapping between a certain vocal tract state and a certain sensory target, such as a vowel (Niziolek and Guenther, 2013). When attempting to produce the specific vowel, the participant initiates a motor sequence and compares their expectations of the intended sound to auditory feedback (Houde and Nagarajan, 2011), which may lead to swift articulatory adjustments if the vowel is off-target (Niziolek et al., 2013). During the identification task, the participant is exposed to repeated ambiguous and unambiguous stimuli produced by the same voice and must map the incoming speech sounds onto abstract phonetic categories in order to respond. This may involve comparing the exposure voice to representations stored in memory, and adjusting to the idiosyncrasies of the exposure voice (Liu and Holt, 2015).

Acoustic-to-phonetic category mapping proceeds normally during the unaltered feedback session, producing a certain proportion of "she" responses in response to changes in the frequency of the fricative while the context remains relatively constant (though repeated presentation or production of a sound may also alter perception; Eimas and Corbit, 1973; Shiller et al., 2009). Introducing the altered feedback during the production task alters the relationship between a given articulatory trajectory and its acoustic outcome, and adaptation reflects a stabilization of this shifted mapping (Purcell and Munhall, 2006). The change in the proportion of "she" responses indicates that shifting this mapping alters perception as well, suggesting that the listeners actively use knowledge of acoustic-to-motor mappings to classify contextually dependent speech sounds.

This experiment contributes to the growing body of research demonstrating that sensory adaptation in speech can influence motoric learning (Bradlow et al., 1997; Lametti et al., 2014a), and that motoric adaptation can affect perception of other speakers (Lametti

et al., 2014b; Shiller et al., 2009). What is most interesting is the pattern of generalization, that articulatory remapping during vowel production affects *CFC* processes involved during consonant perception. Perceptual retuning is notoriously specific (Kraljic and Samuel, 2005; Reinisch et al., 2014). This is somewhat true as well for adaptation to altered feedback. Thus far, generalization of articulatory adaptations has only been examined in with regard to generalization in production changes; such studies have found that changes in articulation are not limited to the adapted segment, but can also generalize to other tone categories (Jones and Munhall, 2005), or other vowels (Cai et al., 2010). Our results demonstrate that perceptual changes can also occur even if the adapted sound is only contextually related to the target sound, as in the case of *CFC*. Testing the effects of adaptation on perceptual compensation for coarticulation eliminated the possibility that the adaptation-induced reported here and in previous experiments (Lametti et al., 2014b; Shiller et al., 2009) can be attributed to response biases. Finally, the experiment capitalized on the fact that when participants oppose a shift in auditory feedback, the opposing response is not complete (Katseff et al., 2012). Thus, participants with differing articulatory responses heard very similar auditory feedback, enabling adjudication between articulatory and acoustic accounts for the observed effects.

While perception may not depend on production, it is clear from this study and others that sensorimotor processes can and do affect perception. Such a view accords with models of speech perception that suggest that the motor system, rather than playing a crucial role in the online decoding of speech sounds, plays a more modulatory but still important role in sensorimotor integration during speech perception (Hickok et al., 2011). Accurately identifying rapidly coarticulated segments is a common problem listeners must face in decoding a continuous speech signal. Listeners have a wealth of production experience available to them, and our results, as well as others, suggest that this experience is deployed online to assist in decoding speech (Poeppel et al., 2008). Sensorimotor integration processes may serve to increase the intelligibility of accented speakers (Adank et al., 2010, 2013) or support phonetic alignment between interlocutors (Pardo, 2006). Exploring modulatory relationships such as these may help to reconcile disparate bodies of research that focus on production and perception in isolation.

Chapter 3

Mapping the Speech Code: Cortical responses linking the perception and production of vowels.

The acoustic realization of speech is constrained by the physical mechanisms by which it is produced. Yet for speech perception, the degree to which listeners utilize experience derived from speech production has long been debated. In the present study, we examined how sensorimotor adaptation during production may affect perception, and how this relationship may be reflected in early vs. late electrophysiological responses. Participants first performed a baseline speech production task, followed by a vowel categorization task during which EEG responses were recorded. In a subsequent speech production task, half the participants received shifted auditory feedback, leading most to alter their articulations. This was followed by a second, post-training vowel categorization task. We compared changes in vowel production to both behavioral and electrophysiological changes in vowel perception. No differences in phonetic categorization were observed between groups receiving altered or unaltered feedback. However, exploratory analyses revealed correlations between vocal motor behavior and phonetic categorization. EEG analyses revealed differences between groups, as well as correlations between vocal motor behavior and cortical responses in both early and late time windows. These results suggest that participants' recent production behavior influenced subsequent vowel perception. We suggest that the change in perception can be best characterized as a mapping of acoustics onto articulation.

3.1 Introduction

Learning to produce speech requires mapping acoustics onto articulation (Guenther, 1994; Kuhl, 2004). Sensory-to-motor mappings may be continuously updated during adulthood based on input from the environment (e.g., Sancier and Fowler, 1997) and sensorimotor experience (Brainard and Doupe, 2000; Tschida and Mooney, 2012). While the role of sensorimotor experience for maintaining production abilities is uncontroversial, the role of sensorimotor experience during speech perception has been highly contested (Hickok, 2009; Hickok et al., 2009; Wilson, 2009). More recently, the focus has shifted from investigating whether production systems are involved in perception to 'unpacking' how production systems and production experience influence perception (e.g. Skipper et al., 2017; Stasenko et al., 2013). While sensory-to-motor mappings appear to be critical for developing speech production abilities, it is unclear to what perception may involve mapping acoustics onto articulation. Evidence suggests that more accurate speech perception correlates with more distinct articulation (Perkell et al., 2004a,b), pointing towards a close link between perception and production abilities. Yet it is not clear how changes in one system (e.g., perception) lead to changes in the other (e.g., production). In the present study, we examined cortical and behavioral responses during a vowel categorization task prior to and following sensorimotor training in order to investigate how sensorimotor experience affects the neural processing of speech sounds.

Phonetic categories that can differ by a single acoustic value, such as voice-onset-time, are divided by a perceptual boundary (Liberman et al., 1957), marking the point at which sound acoustics stop corresponding to one category and begin to correspond to the other. The location of this perceptual boundary along a continuum between two sound categories can be shifted as a result of experience, a phenomenon known as phonetic recalibration (Samuel and Kraljic, 2009). For example, by inserting an ambiguous fricative sound between [f] and [s] into a context in which hearing the sound as [s] would create a real word and [f] would not (e.g., pass vs. paff), listeners can be biased to perceive the sound as [s]. After repeated exposure to these biasing contexts, listeners are more likely to categorize the ambiguous sound as [s] in subsequent phonetic categorization tasks, a process known as 'phonetic recalibration' (Norris et al., 2003). Thus, experience that biases how acoustic values are a categorized can lead to shifts in the perceptual boundary between two phonetic categories.

Recent experiments have found that sensorimotor adaptation can also lead to shifts in the perceptual boundary between two phonetic categories (Lametti et al., 2014b; Shiller

et al., 2009). Frequency alteration devices (e.g., Houde and Jordan, 1998, 2002) enable an experimenter to introduce a mismatch between a speaker's articulation and the acoustics of the resulting sound. A speaker may attempt to compensate for the shift by articulating in the opposite direction, though the degree of compensation is usually not sufficient to completely counteract the frequency shift (Katseff et al., 2012; MacDonald et al., 2011). After continued exposure to shifted feedback, the compensatory response may stabilize such that when producing a target sound, the speaker continues to utilize a newly learned articulation even when the altered feedback is masked or removed (Purcell and Munhall, 2006). At this point of stabilization, the speaker is considered to have 'adapted' to the new sensorimotor mapping. Shiller et al. (2009) found that when participants' [s] productions were shifted down (towards values for [ʃ]), participants compensated by increasing the frequency of the fricative. Compared to baseline, this change in production behavior led participants to categorize more stimuli as [s] following training. In contrast, control participants who received unaltered feedback tended to categorize *less* stimuli as [s] after training.

In a related study, Lametti et al. (2014b) found that changes in vowel articulation due to sensorimotor adaptation led to specific changes in phonetic categorization. Participants were first tested on their perception of a phonetic continuum between 'head' and 'hid' (Exp. 1) or 'head' and 'had' (Exp. 2). Then, during production training, participants produced the word 'head', while F1 was either increased (to sound more like 'had') or decreased (to sound more like 'hid'). Following sensorimotor adaptation, participants who *articulated* into the test region (e.g., producing 'head' more like 'hid', then tested on a head-to-hid continuum) were found to show a decrease in the proportion of stimuli labeled as 'head'. No changes in categorization were found for the opposite shift or control participants. While neither study found significant correlations between the magnitude of adaptation and the magnitude of change in perceptual function, such studies demonstrate that sensorimotor adaptation can lead to changes in phonetic categorization.

However, it is unknown whether the effects observed in these experiments stem from changes to early stages of speech sound processing, e.g., acoustic encoding or feature extraction, or later stages involving perceptual decision making (Mostert et al., 2015). This distinction is crucial in order to relate these effects to speech perception under typical listening conditions, as changes to late stage processes may only affect performance on specific laboratory tasks (Hickok and Poeppel, 2000, 2007). Furthermore, examining how a listener's sensorimotor experiences alter the processing of sounds may elucidate the role of sensorimotor integration in speech perception (Hickok et al., 2011).

We consider two primary time-windows at which sensorimotor experience may affect speech sound processing. The first is an early window around 100ms after stimulus

onset, corresponding to the N1/M1 electrophysiological components. The N1/M1 has been described as an 'exogenous' response (Picton, 2013), reflecting the acoustic properties of the stimulus. Accordingly, repeated presentation of a speech stimulus leads to suppression of this component, while actively imagining the same stimulus prior to presentation does not (Tian and Poeppel, 2013).

The identity of a perceived vowel can be predicted based on early tonotopic activity in primary auditory cortex (Chang et al., 2010) that encodes the acoustic features relevant for distinguishing vowels from each other. Many vowels can be described as a combination of the first two resonating frequencies, or *formants*, of the vocal tract (F1 and F2). The values of these formants correspond to the height of the jaw and tongue body (F1) and the anteriority/posteriority of the tongue body (F2) (Fant, 1960). Vowels varying along these two dimensions elicit distinct cortical responses as early as 100ms after stimulus onset (Obleser et al., 2003a,b, 2004; Shestakova et al., 2004). Based on these data, early auditory activity around 100ms may reflect acoustic feature extraction (Tavabi et al., 2007) or pre-lexical abstraction (Obleser and Eisner, 2009).

While long term changes in the amplitude of the N1/M1 auditory component have been found after musical training (Pantev et al., 1998), the amplitude of activity in this time-region can also be modulated by attention (Poeppel et al., 1997). Hickok et al. (2011) have speculated that forward predictions based on prior sensorimotor experience direct attention to relevant acoustic features of an expected sound, possibly modulating the gain and response selectivity of neurons tuned to those features. If sensorimotor experience can affect how features are extracted or encoded, e.g., by altering the degree of vowel 'height' encoded by a particular F1 value or the degree of vowel 'frontness' encoded by F2, then this ought to be reflected by changes in N1 amplitude.

The second time window we consider is centered around 200ms (P2/M2) and has been associated with perceptual decision making (Mostert et al., 2015) as well as phonological processing (Tian and Poeppel, 2013). While it may be possible to decode vowel identity from distributed activity in early processing stages (Chang et al., 2010), in a phonetic categorization task this neuronal activity must ultimately be linked to a linguistic representation in order to produce a behavioral response (Poeppel et al., 2008). Phonetic categorization involves mapping a stimulus exemplar drawn from a continuous acoustic distribution onto a discrete category. Results from such experiments give rise to a sigmoidal response curve, marking the boundary between the two phonetic categories (as, for example, in the present experiment; see Fig. 3.2). Due the transform from a continuous acoustic space to a binary response space, behavior in sensory decision tasks may not directly reflect sensory encoding but subsequent decision processes *acting upon* sensory representations (Mostert et al., 2015). Accordingly, behavioral responses

in a phonetic categorization task have been found to correlate with variations in the amplitude of the event-related P2 component (but not the earlier N1; Bidelman et al., 2013).

Auditory training with speech stimuli, in which participants respond to training stimuli with non-vocal responses (e.g., button presses), has been found to modulate P2 amplitude. The effects of auditory training on cortical responses has been investigated extensively with regard to the perceptual learning of voice-onset-time (VOT) contrasts (Tremblay et al., 1997, 2001). This series of auditory training and auditory exposure studies revealed that P2 amplitude increases in response to repeated exposure to a training continuum, regardless of change in perceptual performance (Alain et al., 2010; Sheehan et al., 2005; Tremblay et al., 2010, 2009). Researchers have consequently suggested that increases in P2 amplitude are general biomarkers of auditory learning, possibly representing a first-stage process involving auditory object familiarization and representation (Tremblay et al., 2014). However, the authors specifically ascribe this increased P2 amplitude to the context of learning a *novel* (i.e., unknown, non-native) phonetic contrast representing a distinct auditory object (Ross et al., 2013). Therefore it is unclear from these studies whether changes to *existing* phonetic contrasts also involves modulation of this component.

Electrophysiological experiments on native speech categories suggest that phonetic recalibration results in changes to later perceptual decision components. Utilizing a mismatch negativity paradigm, van Linden et al. (2007) exposed participants to an ambiguous consonant midway between [t] and [p]. This ambiguous stimulus was utilized as the standard, and compared with a deviant which was an unambiguous [t]. By altering the lexical context in which this ambiguous stimulus was embedded, listeners were biased to hear the standard as either a [p] or a [t], which in previous experiments elicited a shift in the phonetic categorization boundary of a stimulus continuum between [p] and [t] (van Linden and Vroomen, 2007). A significant mismatch negativity was elicited when the standard was heard as [p] but not when it has been heard as [t], suggesting that biasing the listeners to categorize the ambiguous stimulus as a member of another phonetic category resulted in greater perceptual distance between standard and deviant. Furthermore, the peak MMN response was found at 215ms after segment onset, approximately the same time-region implicated in auditory perceptual learning (Tremblay et al., 2014). If the processes involved in phonetic recalibration are similar for both sensory (Samuel and Kraljic, 2009) and sensorimotor (Lametti et al., 2014b) training, then we may also expect sensorimotor training to elicit changes in cortical amplitude in this late time window.

Ito et al. (2016) examined the effects of sensorimotor adaptation on auditory potentials recorded in response to a single unambiguous [ɛ] vowel, which was presented before and after speech motor training. This motor training involved shifting auditory feedback during production of the word "head" such that the participants heard themselves producing a vowel more like the one in "hid" (by decreasing F1). In order to counteract this shift in feedback, participants would therefore have to produce a vowel more like that in "had" (by increasing F1). Participants were divided into three groups of equal size based on whether they had produced consistent compensatory motor behavior opposing the feedback shift (adapted), had failed to compensate for the shifted feedback (non-adapted), or had instead received unaltered feedback (control). Only in the adapted group did the authors find a significant change in the amplitude of the P2 component. In contrast to the increases in P2 amplitude found in perceptual learning studies (Tremblay et al., 2010, 2014), adapters exhibited a *decrease* in P2 amplitude over right frontal electrodes. While the interpretation of the decreased P2 amplitude was not entirely clear, the timing of the effect was in line with previous research on phonetic recalibration and perceptual learning (Tremblay et al., 2014; van Linden et al., 2007).

To summarize, previous research has found that sensorimotor adaptation to altered auditory feedback during speech production alters phonetic categorization (Lametti et al., 2014b). The latency of sensorimotor adaptation effects on cortical responses (Ito et al., 2016) suggests that sensorimotor adaptation modulates activity in processing stages associated with phonetic categorization (Bidelman et al., 2013) rather than acoustic encoding (Obleser and Eisner, 2009).

The present study sought to build upon these results in order to further explore how speech perception may reflect sensorimotor experience. We compared behavioral and cortical responses during a phonetic categorization task prior to and following sensorimotor training of speech production. Dutch participants were recorded producing the Dutch word "pet" ("cap") containing the front mid-vowel [ɛ], and then performed a phonetic categorization task during which EEG was recorded. For the categorization task, we parametrically varied values of F1 to create a five-step continuum between [ɛ] and [ɪ]. In the subsequent speech training session, half the participants were exposed to altered auditory feedback (the AF group) while the other half received unaltered feedback (the UF group). For the AF group, the value of F1 was increased, which caused participants to hear themselves producing a vowel more like [æ]. Compensating for this shifted feedback required articulating into a motor space that would normally produce a sound between [ɛ] and [ɪ], which in previous experiments had been found to lead to phonetic recalibration (Lametti et al., 2014b). This training session was followed by another phonetic categorization task (Fig. 3.1). The design enabled us to examine how changes in phonetic recalibration were related to changes in speech motor behavior,

and how changes in speech motor behavior and phonetic categorization were reflected in electrophysiological responses.

In (Lametti et al., 2014b), articulating [ɛ] as a more [ɪ]-like vowel led to increases in the proportion of stimuli categorized as [ɪ]. We therefore expected that adapters (AF group) would categorize more stimuli as [ɪ] after sensorimotor adaptation, while controls (UF group) would not. However, while we observed significant adaptation in response to the altered feedback, we found no significant differences between groups in changes in phonetic categorization after sensorimotor training.

We therefore conducted systematic exploratory analyses to examine to what extent individual differences in the production of the training vowel ([ɛ]) corresponded to behavior in the phonetic categorization tasks and electrophysiological data (cf. Bradlow et al., 1996). Though no significant correlations between perceptual and motoric behavior were found in Lametti et al. (2014b), based on their group level results we expected that decreases in F1 (articulating an [ɛ] as a more [ɪ]-like vowel) should correlate with an increase in the proportion of stimuli categorized as [ɪ], though possibly only for participants who received altered auditory feedback.

Regarding the electrophysiological data, if auditory-motor remappings lead to changes in vowel encoding, then this should modulate the amplitude of the early N1 component. If, auditory-motor remappings cause changes in perceptual decision making, then this ought to modulate the later P2. With respect to the direction of this modulation, prior evidence leads to conflicting predictions. While exposure to a phonetic continuum leads to increases in P2 amplitude (e.g., Tremblay et al., 2014), sensorimotor adaptation has been found to lead to decreases in P2 amplitude (Ito et al., 2016). However, the fact that the sensorimotor adaptation used only a single test vowel, rather than a continuum, may have led to this discrepancy in the results. If exposure to a phonetic continnum leads to increases in P2 amplitude, while sensorimotor adaptation leads to decreases in P2 amplitude, we may expect these effects to cancel out for adapters (or go in opposite directions for specific stimuli), while controls should only exhibit increases in P2 amplitude.

Rather than basing our expectations about changes to neural components based on the type of feedback participants receive, alternatively, we can generate predictions about the direction of neural component change based on what would be expected from a sensory-to-motor mapping. Bidelman et al. (2013) found that an ambiguous vowel between [a] and [u] was classified as [u] (closed jaw position/low F1), P2 amplitude was greater than when the same vowel was perceived as [a] (open jaw position/high F1). If the vowel [ɛ] in “pet” comes to be associated with a more closed or open jaw position

due to changes in production (regardless of feedback), then a sensory-to-motor mapping account would predict that P2 amplitude should increase or decrease correspondingly. Specifically, if adapters produce [ɛ] with a more closed jaw position, we expect to observe increases in the amplitude of the P2 component in the subsequent perceptual task.

3.2 Material and Methods

3.2.1 Participants

A total of 48 native Dutch speakers took part in the study. All reported normal hearing and vision. Previous experiments had found that some participants do not exhibit changes in articulatory behavior in response to altered auditory feedback (MacDonald et al., 2011). Twenty-eight participants were assigned to the altered feedback condition. Of these, twenty produced significant articulatory responses opposing the direction of the shifted feedback (see Results: speech adaptation; average age = 21.3, range = $18 - 28$, four men). We then recruited an additional twenty participants to serve as a control group (unaltered feedback; average age = 22, range = $19 - 30$, five men).

Ethical approval for this study was obtained from the Ethics Committee of the Social Sciences Faculty of Radboud University. Participants were informed that their participation was voluntary and that they were free to withdraw from the study at any time without any negative repercussions and without needing to specify a reason for withdrawal. Written consent was obtained from each participant and all were reimbursed for their participation.

3.2.2 Procedure

After cap fitting, participants were led to a recording booth for the production baseline. Participants were then led back to the EEG recording booth for the perception pre-test. They then returned to the recording booth for production training, and then once again returned to the EEG booth and performed the listening post-test (which was identical in design to the pre-test). In order to ensure that listening effects in the post-training phase were based solely on the feedback received during training and not from vicarious inter-session speech, all participants were instructed to not communicate verbally (unless absolutely needed) between the training phase and the listening post-test, while the researcher also refrained from any verbal communication as well.

3.2.3 Speaking - baseline and training

Production tasks took place in a sound-attenuated booth. Speech recording and auditory feedback transmission was carried out using Audapter (Cai et al., 2008; Tourville et al., 2013), a feedback manipulation program implemented in Matlab (Mathworks, 2012). Participants were seated in a chair approximately 5 to 10cm away from a pop-filter shielded microphone (Sennheiser ME 64), and fitted with sound isolating headphones (Sennheiser 280). The volume of the headphones was calibrated individually such that participants reported only being able to hear their voice through the headphones, masking their actual productions. Yet we also ensured that the level of the volume caused no physical discomfort. If participants began to whisper or speak too softly, an automated warning message appeared on screen asking them to increase their speaking volume. Broadband noise (60dB) was added to the auditory feedback signal in order to further mask bone-conducted sound (Békésy, 1949).

Speech tokens were elicited by visual presentation of the orthographic form of the target word. In the baseline phase, participants first produced four instances of the words "pit" (pit) and "pet" (cap/hat; containing the vowels [ɪ] and [ɛ], respectively) in random order. They then produced 50 repetitions of word "pet". Averaged F1 measurements in these 50 trials constituted each participant's production baseline.

In the training phase, participants repeated the Dutch word "pet" ("cap") 100 times. For participants in the control group, there were no modifications to the spectral parameters of the participants' utterances (Unaltered Feedback; UF). However, the auditory signal was transmitted through the same speech modulation software in order to generate the same delay and masking noise experienced by the altered feedback group. For the altered feedback (AF) group, we implemented a slightly modified version of the paradigm utilized in Lametti et al. (2014b). For this group, the frequency of the first formant was shifted upwards by 25%. With this increase in F1, if a participant produced the word "pet" ([pɛt]) with normal articulation, at the maximum value of the feedback shift the participant would hear themselves producing something like the English word "pat" ([pæt]). The intensity of the feedback shift increased linearly from between trials 1 through 30 and was held constant for the remainder of the training session.

3.2.4 Listening - EEG data acquisition and preprocessing

EEG data acquisition took place during the two listening tasks following the production baseline and training sessions. Participants were seated comfortably in front of a computer screen and a button-box. Auditory stimuli were emitted from two speakers

flanking the computer screen. Stimulus delivery and response monitoring was controlled using Presentation (Version 0.70, www.neurobs.com). Each trial began with a blank screen. Auditory stimulus presentation began after a random waiting interval between 400 and 600ms (in increments of 20ms; Bidelman et al., 2013). Participants attempted to respond as quickly as possible by pressing one of two buttons corresponding to two phonetic categories: "korte e" ("short e", [ɛ]) and "korte i" ("short i", [ɪ]). All participants responded using the index and middle fingers of their dominant hand.

Auditory stimuli comprised a five-step Klatt-synthesized vowel continuum between clear [ɛ] and clear [ɪ]. Pitch and formant values were based on average values for a female speaker of Dutch (Schuerman et al., 2015). Stimuli were presented 100 times each. Sessions thus consisted of a total of 500 trials, with self-paced rest periods after every 100 trials (five blocks). Presentation was randomized such that an equal number of presentations of each stimulus (20) occurred in each block, and every possible two-way combination of stimuli (e.g., step 1 – step 1, step 1 – step 2...) occurred an equal number of times.

Continuous electroencephalograms (EEGs) were recorded using a 32 electrode Acti-Cap system, with reference electrodes placed on both mastoids. Two additional electrodes were placed above and below the left eye to measure blinks and eye-movements. A common ground electrode was placed along the midline (AFz), with reference electrode on the left mastoid. EEG data was later re-referenced to paired electrodes on both left and right mastoids.

EEGs were sampled at 20 kHz and online filtered between 0.05 and 3500 Hz. Artifact rejection and averaging was conducted in Matlab (Mathworks, 2012) using the Fieldtrip toolbox (Oostenveld et al., 2011). Event-related potentials (ERPs) were baselined with respect to -100ms prior to stimulus onset and windowed from -100ms pre-stimulus onset to 600ms post-stimulus onset. Prior to artifact rejection, a band-stop filter with a 50Hz center frequency and 1Hz bandwidth was applied to eliminate machine noise. Semi-automated artifact rejection was employed, in which likely artifacts were marked. All trials containing artifacts were removed after visual inspection. This resulted in a total of 4227 rejected trials (22.1% of data; average 52.2 trials per participant). For each participant, independent component analysis was used to identify and remove eye-blink and heartbeat related components. Prior to averaging, the remaining ERPs were low-pass filtered at 30Hz to isolate cortical responses.

3.2.5 Acoustic analysis

Recordings from the production task were analyzed with Praat (Boersma and Weenink, 2016). The vocalic section of each recording was automatically located, and vowel measurements taken from the midpoint. The values of the first and second resonant frequencies of the vocal tract (formants; F1 and F2) were calculated using 10ms overlapping windows, and tracked using 12 LPC coefficients. Formant values exceeding five standard deviations above or below each participants average F1 value were excluded as these values were likely the result of tracking errors. Formant values in Hertz were converted to Mels, a logarithmic frequency scale based on the properties of human hearing, using the formula $2595 * (log(1 + (F2\_Hz/700)))$.

In order to compare participants having differing vocal tracts with respect to changes in production, formant values were standardized relative to each participant's average values during the baseline speaking task, according to the following equation (where F refers to F1 or F2):

$$F_{standardized} = (F - mean(F_{baseline}))/sd(F)_{baseline} \tag{3.1}$$

3.2.6 Event related potential analysis

Event-related potentials were analyzed using nonparametric cluster-based permutation analysis (Maris and Oostenveld, 2007), which is well suited to exploratory comparisons between two groups. This method utilizes an algorithm based on the assumption that ERP effects are clustered over both space and time in order to address the family-wise error rate arising from multiple comparisons. For each sample, the candidate contrasts are compared at each channel and each time point using t-tests. Next, all samples with t-values larger than a specified threshold are selected, while all samples failing to meet this threshold are discarded. Selected samples are then clustered on the basis of temporal and spatial distance, and t-values are summed over these clusters. In order to generate the null distribution against which this test statistic is compared, trials from the two conditions are randomly partitioned into two subsets and summed t-values are calculated from these randomly generated clusters. In this experiment, the number of permutations per contrast was set at 10000. The test statistics of these randomly generated partitions are compared to the test statistic of the experimentally observed data, and the proportion of partitions greater than the observed partition constitutes the significance level of the cluster (i.e., its *p-value*). In the current experiment, within- and between-group contrasts were tested using a two-tailed alpha level of 0.025. Tests

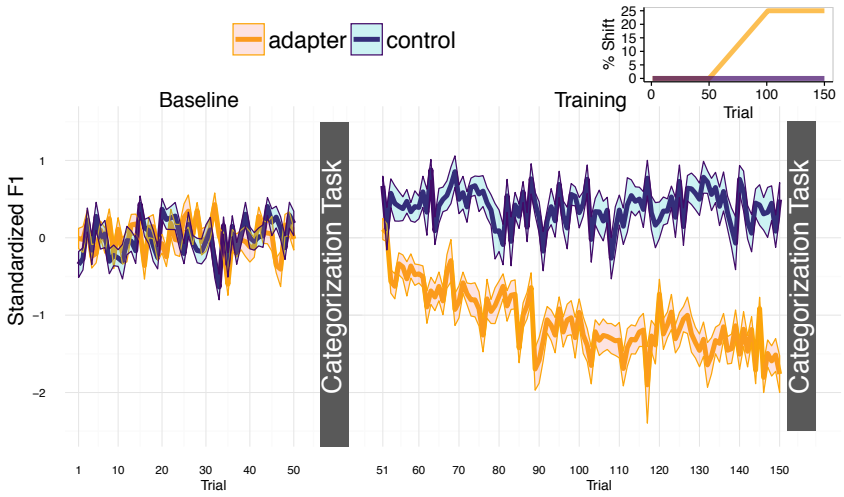


FIGURE 3.1: **Experimental design and production data.** In the two speaking task (baseline and training), participants repeatedly produced the Dutch word "pet" (cap), which contains the front-mid vowel [ɛ]. Phonetic categorization tasks took place immediately after the end of the baseline (trial 50) and training sessions (trial 150). For both groups, feedback was unaltered during the baseline session (left panel). For the altered feedback group, the value of F1 in the auditory feedback increased linearly between trials 51 and 70 to a maximum of 25% greater than each trial's original value (inset, top right). Thus, participants heard themselves producing a more [ɪ]-like vowel, which led to compensatory decreases in F1 (yellow line). Feedback was unaltered for controls (purple line).

for significant clusters in the interaction between session and group utilized an alpha of 0.05.

3.3 Results

All statistical analyses were implemented in R (R Development Core Team, 2013). ANOVAs were implemented using the *ez* package (Lawrence, 2011). Bayesian ANOVAs were calculated the package *BayesFactor* (Morey et al., 2015) and Bayesian correlations were conducted using the package *BayesMed* (Nuijten et al., 2015). In cases where the distribution of the data did not allow for parametric testing, the appropriate non-parametric version was utilized.

3.3.1 Speech production training

For participants in the altered feedback group, adaptation was assessed utilizing one-tailed independent sample t-tests between the baseline (50 trials) and the last 50 trials of

the training phase (hold phase) for each individual. Out of 28 participants, 20 "adapters" exhibited significant compensatory decreases in F1, opposing the shift in auditory feedback. The remaining eight participants were excluded from the subsequent analysis. We first tested for potential phonetic differences between groups during the baseline and training sessions. While the experimental manipulation targeted F1, speakers have been found to alter their production of unshifted formants in response to altered auditory feedback (MacDonald et al., 2011). Therefore, in addition to F1, we also analyzed F2.

An ANOVA on produced F1, with type as a between-subjects factor and session as a within-subjects factor, revealed a significant main effect of session ($F(38) = 5.676$, $p = 0.022, ges = 0.011$), as well as a significant interaction between session and type ($F(38) = 41.448, p < 0.001, ges = 0.077$). Subsequent two-sample t-tests (with Levene tests for equal variance) indicated that F1 differed significantly between groups in the training session ($t(38) = 3.618, p < 0.001, BF_{10} = 35.26$), but not in the baseline session ($t(38) = 0.021, p = 0.98, BF_{10} = 0.31$). Thus, groups did not differ on F1 during baseline but did differ significantly during the training session.

For F2, only a significant interaction between session and type was found ($F(38) = 8.523, p = 0.006$). However, the effect size was extremely small ($ges = 0.003$), and post-hoc tests indicated no significant differences in F2 between groups in either the baseline ($p = 0.85, BF_{10} = 0.31$) or training sessions ($p = 0.396, BF_{10} = 0.41$).

In the hold phase of the training session, controls' exhibited an average increase in F1 of 43.69_{mel} relative to baseline, while adapters exhibited an average decrease of -94.54_{mel}. In order to assess changes in formant production between sessions, we standardized formant values with respect to the mean and standard deviation of each participant's baseline. Figure 3.1 displays the average values for standardized F1. Two-sample t-tests confirmed that average standardized F1 differed significantly between adapters and controls ($t(38) = -5.859, p < 0.001, BF_{10} = 13941.8$). A Wilcox test indicated that standardized F1 was significantly below baseline for participants in the AF group ($p < 0.001$). At first, standardized F1 in the control group did not appear to differ significantly from baseline. However, after removing one possible outlier ($z = -2.93$), standardized F1 was found to be significantly greater than baseline in the control group ($t(18) = 3.680, p = 0.002$). Thus, these groups exhibited divergent F1 productions during the speech production tasks.

Standardized F2 differed significantly between adapters and controls as well ($t(38) = 2.4567, p = 0.019, BF_{10} = 3.09$). Standardized F2 differed significantly from baseline for controls ($t(18) = -2.28, p = 0.034$) but not adapters ($p = 0.17$).

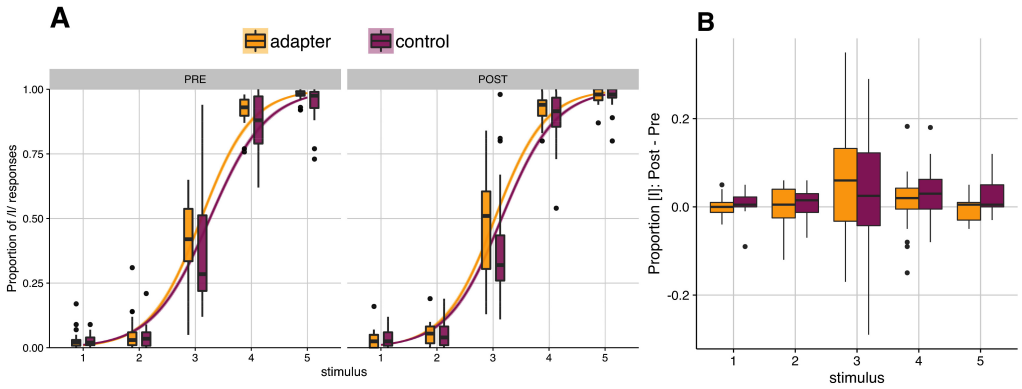


FIGURE 3.2: **Phonetic categorization.** A. Results of phonetic categorization tasks before and after speech production training. After baseline (left panel) and training (right panel) speaking sessions, participants categorized vocalic stimuli as either [ɛ] or [ɪ]. The proportion of [ɪ] responses is indicated on the y-axis, as a function of stimulus step (x-axis). Yellow (adapters) and purple (controls) lines indicate the slope of a logistic function fit to individual responses. Box plots indicate distribution of results across participants. No significant differences were found between groups in either session. B. Box plots of average change in the proportion of stimuli categorized as [ɪ] by stimulus. Most participants tended to categorize all stimulus steps more often as [ɪ] in the post-test session. Outliers indicated by filled circles.

In the control group, Subsequent two-tailed t-tests for each participant revealed that nine participants exhibited statistically significant increases in F1 compared to baseline, while two exhibited significance decreases. Similarly, eight controls and five adapters exhibited significant decreases in standardized F2, while three controls and eight adapters exhibited significant increases in standardized F2. Therefore, both within and across groups, participants exhibited variable patterns of vocal motor behavior, not only for the altered formant but in other formants as well.

3.3.2 Phonetic categorization

Prior to speech production training, control participants were found to have categorized slightly fewer stimuli as [ɪ] than adapters (3.2A, left panel). Following training, both groups categorized more stimuli as [ɪ] (Fig. 3.2A, right panel). The average proportion of stimuli categorized as [ɪ] was analyzed using repeated-measures ANOVA (with Greenhouse-Geisser corrections), with group (adapters/controls) as a between-subjects variable and stimulus (five levels), session (two levels), and block (five levels) as within-subjects variables. Significant main effects were found for stimulus ($F(4) = 883.29$, $p < 0.001, ges = 0.898$) and block ($F(4) = 3.741, p = 0.01, ges = 0.007$). The main effect of block, with no interaction, indicated that for both groups, the proportion of [ɪ] responses tended to increase over both sessions (Table 3.1). A significant interaction was also found between stimulus and session ($F(4) = 3.379, p = 0.044, ges = 0.004$),

TABLE 3.1: **Average proportion of [ɪ] responses by block, session, and group.**

Session			PRE					POST		
Block	1	2	3	4	5	1	2	3	4	5
Adapters	0.46	0.47	0.5	0.48	0.49	0.48	0.50	0.52	0.49	0.48
Controls	0.43	0.43	0.44	0.48	0.46	0.46	0.46	0.47	0.49	0.47

though as this effect did not pertain to our hypothesis it was excluded from further investigation. Unlike in Lametti et al. (2014b), there was no significant effect of group and no interaction between group and session (all $ps >= 0.116$).

While between-group differences failed to reach significance, participants in both groups exhibited a large amount of variation in behavioral responses (Fig. 3.2A–C). This variation between participants (in both groups) was the object of our exploratory analyses. We specified several potential relationships of interest between vocal motor behavior (quantified as F1 and F2) and phonetic categorization. These included: 1) Correlations between average F1 and F2 values in the baseline production task and average proportion of [ɪ]-responses in the subsequent perception task; 2) Correlations between average F1 and F2 values in the hold phase of the training session and average proportion of [ɪ]-responses in the subsequent perception task; 3) Correlations between standardized F1 and F2 (representing the change in formant values with respect to each participants baseline) and the between session difference in average proportion of [ɪ]-responses. This totaled six correlations.

While none of these correlations was found to be significant after applying Holm–Bonferroni corrections for multiple testing (Holm, 1979), Bayesian analyses suggested some evidence of possible correlations between vocal motor behavior and perceptual responses. We therefore report the correlations with Bayes factors and the corresponding non-significant p-values.

Due to the presence of significant correlations between F1 and F2 in both sessions, we conducted partial correlations controlling for the value of the other formant. Correlation tests between average [ɪ]-responses reported in the pre-training session and average F2 produced during baseline suggested a potential relationship (Fig. 3.3A; $r = -0.37$, $p = 0.022, BF_{10} = 2.42$). The correlation between average F2 in the training session and the proportion of [ɪ] responses post-training did not reach significance ($r = -0.27$, $p = 0.096, BF_{10} = 0.72$). Our tests did not suggest, in either session, any relationship between average proportion of [ɪ] responses and F1.

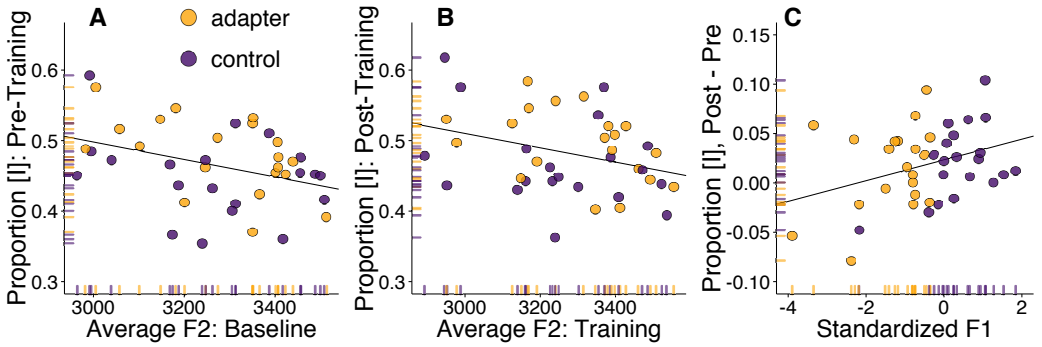


FIGURE 3.3: **Correlation analyses between speech motor behavior and phonetic identification.** Adapters are shown in yellow, controls in purple. Dotted lines indicate slope of regression line. Black lines and text represents correlation across groups, while colored lines represent within-group correlations. **A**: Correlation between F2 (in mels; averaged over all stimulus steps) for the word "pet" produced in the baseline session and the proportion of stimuli categorized as [ɪ] in the following phonetic categorization task. **B**: Correlation between F2 in the speech training session and post-training phonetic categorization. **C**: Correlation between standardized F1 (representing change in F1 in the speech training session compared to baseline) and the difference in phonetic categorization before and after speech training.

A potential relationship was also found between standardized F1, indicating how F1 changed in the training session with respect to each participant's baseline, and between-session changes in the proportion of stimuli categorized as [ɪ] (Fig. 3.3C; $r = 0.34$, $p = 0.034, BF_{10} = 1.17$). The direction of the correlation indicates that participants who produced vowels with a lower F1 during training (primarily adapters), tended to categorize fewer stimuli as [ɪ], while participants who produced vowels with a higher F1 tended to categorize more stimuli as [ɪ]. These results run contrary to those of Lametti et al. (2014b), in which similar adaptation was found to correspond to a group-level increase in [ɪ]-responses.

Follow-up within-group exploratory tests suggested no group-specific correlations. No correlations with standardized F2 were found.

These analyses suggest that overall differences in phonetic categorization may have been related to differences in vocal motor behavior, though unexpectedly, the primary locus of these individual differences was found in variation in F2, not F1. Conversely, between-session changes in phonetic categorization were potentially reflected in standardized F1 but not standardized F2. This may have been due to the greater amount of variation in this formant's value between sessions compared to standardized F2. In both cases, these correlations suggest that the participants' perception of the phonetic continuum may have been influenced by their immediately preceding vocal motor behavior.

3.3.3 Event related potentials

Having observed evidence of potential relationships between the production and categorization tasks, we examined whether similar relationships may be found between behavior in the production tasks and electrophysiological responses. Auditory potentials to each stimulus step, averaged over all participants in the pre-test session, are shown in Figure 3.4C. In contrast to the results of Bidelman et al. (2013), stimulus steps two and three elicited greater P2 amplitudes than stimulus one. This suggests that for the [ɛ] - [ɪ] continuum utilized in this study (Fig. 3.4B), P2 amplitude may not directly reflect distinctions in vowel height.

In the behavioral results, we found evidence suggesting that standardized F1 (representing the change in F1 between sessions) and changes in the proportion of stimuli categorized as [ɪ] may have been related (Fig 3.3C). We also found a potential correlation between the F2 produced by the participants and phonetic categorization in the baseline session (Fig 3.3A). These correlations suggest that the perceptual processing of the vocalic stimuli may have been related to how the participants produced the target vowel during baseline and training. If true, then this predicts that production variables may be reflected in cortical responses recorded during vowel categorization.

We first examined whether standardized F1 was related to either average or stimulus-specific cortical amplitude. In order to do so, it was necessary to identify specific electrodes and time windows over which to average cortical activity. We began by conducting an omnibus ANOVA, testing for main effects of group or interactions between group

TABLE 3.2: **Results of cluster-based permutation analyses on within-group, between-session effects.** All clusters were calculated from stimulus onset to 600ms post-onset. Significant clusters indicated by their time windows (in brackets), with corresponding p-values. Dashes indicate that no positive or negative clusters were observed in the data.

Stimulus	Controls	Adapters
One	[132 - 312] p = 0.022	[92 - 344] p = 0.004
		[396 - 558] p = 0.042
Two	[100 - 202] p = 0.03,	[360 - 586] p = 0.017
	[230 - 336] p = 0.023	
Three	[046 - 336] p = 0.002	—
Four	[110 - 402] p = 0.008	[146 - 358] p = 0.016
		[476 - 600] p = 0.042
Five	[212 - 380] p = 0.012	[090 - 338] p = 0.0041
		[336 - 600] p = 0.006
		[136 - 600] p = 0.044

and session over specific electrodes. No significant effects of group or interactions with group were found.

We therefore decided to investigate stimulus-specific differences in cortical responses using cluster-based permutation analyses (Maris and Oostenveld, 2007). For between-groups contrasts, we found no significant clusters in the pre-training or post-training sessions for any stimulus step. This mirrors the results of the phonetic categorization task, in which no group level differences were found prior to or following speech training. However, within both the adapter and control groups, significant differences were found for between-session contrasts. The results of these within-group analyses are summarized in Table 3.2. For adapters, cortical responses to the endpoint stimuli (one and five) differed most between sessions, while for controls the greatest differences were observed for the most ambiguous stimulus steps (two, three, and four).

In order to determine which electrodes and time-windows to average over in order to compare ERP component amplitudes with behavioral measures, we tested for time-points and electrodes over which between-session activity differed significantly between adapters and controls. This was accomplished by first subtracting averages ERPs in the post-training session from those in the pre-training session for each group and stimulus step. We then performed cluster analyses, comparing these difference waves between groups (3.4 D and E, top and middle panels). No significant clusters were found for between-group differences for any stimulus step when utilizing a time-window from zero to 600ms after stimulus onset. However, visual inspection of between-group difference waves for each stimulus step revealed three possible regions of interest in stimulus steps one and three (Fig. 3.4 D and E, bottom panels, shaded areas), in which the difference in average amplitude exceeded one microvolt. Two of these regions corresponded (approximately) to the N1 and P2 components, though at slightly later time-windows than observed in the grand average ERPs (Fig. 3.4C). As the third region of interest did not pertain to our hypotheses, we excluded it from analysis. Following Bidelman et al. (2013), we selected 20ms time-windows centered around the peaks of these regions of interest, and performed cluster-based permutation analyses averaging over these time windows to test for significane. Significant interactions were indeed found for stimulus one ("P2", 200–220ms, $p = 0.001$) and stimulus three ("N1b", 126–146, $p = 0.01$).

Having identified candidate electrodes and time windows in which adapters and controls differed with regard to between-session variation in cortical amplitude, we then tested whether changes in the amplitude of these two components (averaged over all samples in the 20ms time window and all electrodes found to be active in the cluster) correlated with changes in standardized F1 for these two stimuli. For these four comparisons (two stimuli $*$ two components), we utilized a Bonferroni-corrected p-value of 0.0125.

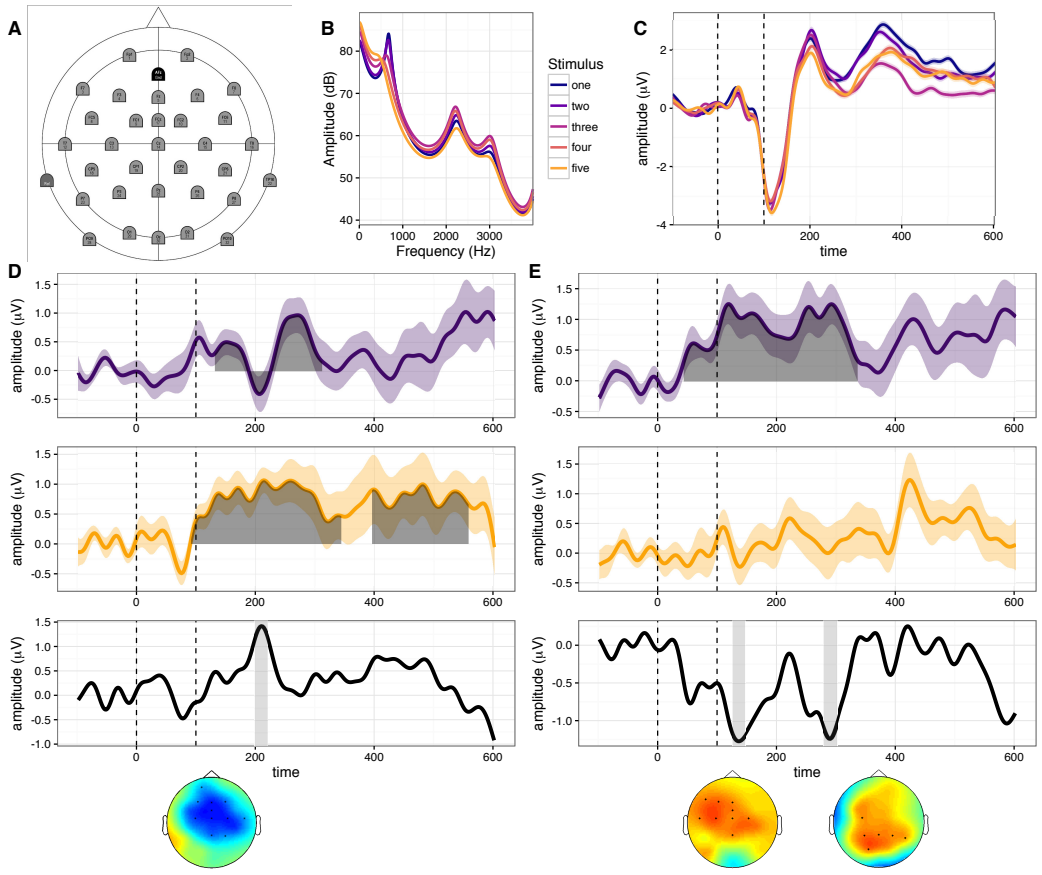


FIGURE 3.4: **Overview of EEG acquisition and results. A.** Electrode layout for EEG acquisition. **B.** Frequency-power spectrum of five stimulus steps ranging from [ɛ] (step one) to [ɪ] (step five). **C.** Cortical responses to each stimulus step during the pre-training phonetic categorization task, averaged over all participants. **D.** Results of cluster-based permutation analyses for stimulus step one. Between-session difference waves are shown for controls (purple line) and adapters (yellow line). Shaded region indicates time-window of significant cluster. Black line (bottom-panel) represents the subtraction of these two difference waves, revealing peak differences at approximately 210ms after stimulus onset. A topoplot indicates the electrodes found to be significant when activity is averaged over a 20ms window (grey vertical bar; P2: 200-220ms). **E.** Results of cluster-based permutation analyses for stimulus step three, for controls and adapters, with topographical plots of significant interactions. The two regions of interest (N1b: 126–146ms, N2: 280–300ms) are indicated by gray vertical bars.

For stimulus step one (clear [ɛ]), we found significant correlations between changes in P2 amplitude and standardized F1 (Fig. 3.5A: $rho = -0.44, p = 0.005$). Follow-up testing revealed no significant within-group correlations. The direction of the correlation indicated that participants who produced vowels with a lower F1 during the training session tended to show increases in P2 amplitude following training, and vice versa. This result runs counter to that found by Ito et al. (2016), in which greater compensatory adaptation led to greater *decreases* in P2 amplitude. However, it is important to note that the feedback shift in their study was opposite in direction to that utilized in this study.

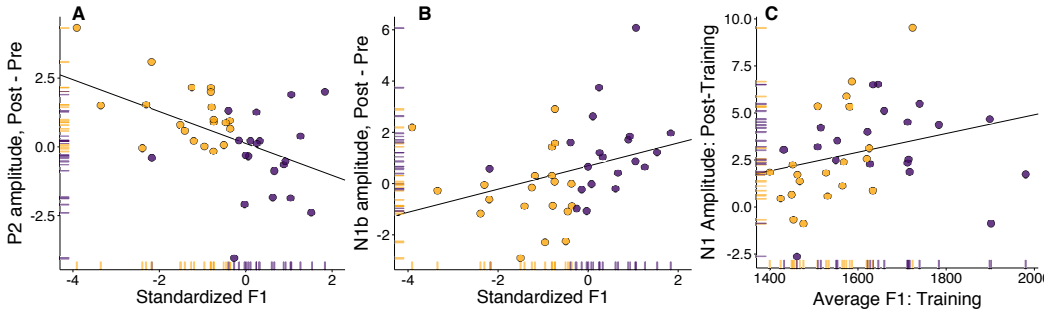


FIGURE 3.5: **Correlation analyses between speech motor behavior and neural component amplitude.** Adapters are shown in yellow, controls in purple. **A**: Correlation between standardized F1 (representing change in F1 in the speech training session compared to baseline) and the difference in P2 amplitude during perception of the trained vowel (step one). **B**: Correlation between standardized F1 (representing change in F1 in the speech training session compared to baseline) and the difference in N1b (126–146ms) amplitude during perception of the most ambiguous vowel (step three). **C**: Correlation between average F1 in the speech training task and P2 amplitude averaged over all stimuli in the post-training identification session.

For stimulus step three (most ambiguous step), changes in "N1b" amplitude were also found to correlate with standardized F1 (Fig. 3.5D: $rho = 0.45, p = 0.004$). These correlations were not significant within either group alone. As most of the between-session changes in perception were localized to this stimulus step (Fig. 3.2C), this might indicate that changes in amplitude of this earlier component drove changes in perception.

Having found some evidence that overall proportion of [ɪ] responses in the baseline session may have been related to produced F2 (Fig. 3.3A), we explored whether F1 and F2 values in the baseline and training session were related to cortical amplitude averaged over all stimulus steps. Utilizing the same electrodes identified by the cluster-based permutation analyses (for early and later components, respectively), we averaged over 20ms time-windows corresponding to peak of the grand average N1 (104–124ms) and P2 (188–208ms) components. Our exploratory tests examined relationships between average component amplitude (N1, P2) and formant production (F1, F2) in both sessions (pre, post), totaling eight comparisons. Partial correlations were utilized to control for the correlated nature of the formants, and Holm-Bonferroni corrections were applied to control for multiple comparisons.

No significant relationships were found, in either session, between average N1 amplitude and formant production. A potential correlation was found between average produced F1 in the training session and average P2 amplitude in the post-training categorization task. The correlation approached significance (Fig. 3.5C; $rho = 0.41, p = 0.01$), and became significant after three multivariate outliers were removed ($r = 0.47, p = 0.004, BF_{10} = 12.2$).

To summarize the results of the neurobehavioral analyses, we found no evidence of within-session, between-group differences in cortical amplitude. However, by-stimulus cluster-analyses suggests that, for specific stimuli, between-session changes in cortical responses varied between adapters and controls. Targeted analyses revealed that for stimulus one, the clear [ɛ] stimulus that was also the training vowel, standardized F1 correlated with changes in P2 amplitude. Yet for stimulus three, the most ambiguous stimulus, changes in standardized F1 were found to correlate with N1b amplitude (slightly later than the grand averaged N1 component). These correlations were not significant within-groups, suggesting that they related to overall changes in produced F1 rather than exposure to altered auditory feedback. The latency of this component suggested that these individual differences in component amplitude reflect changes in acoustic feature processing (Chang et al., 2010; Obleser et al., 2004).

While the behavioral data suggested a possible relationship in between F2 and identification responses, neither N1 or P2 amplitude was found to correlate with produced F2 in either session. The only possible relationship was found between average F2 in the training task and average P2 amplitude in the post-training session.

3.4 Discussion

This study investigated to what extent phonetic categorization reflects one's sensorimotor experience. Primarily, we examined whether perceptual shifts associated with speech production training reflect changes in earlier cortical response components associated with acoustic feature extraction (Chang et al., 2010; Obleser and Eisner, 2009; Tavabi et al., 2007), or later processes associated with phonetic categorization (Bidelman et al., 2013) and perceptual learning (Tremblay et al., 2014). We tested whether altering how a vowel is produced modulates perceptual and electrophysiological responses during phonetic categorization. While between-group differences in phonetic categorization failed to reach significance, we found some evidence evidence linking behavior in the production tasks to behavior in the phonetic categorization tasks. In addition to these relationships between production and perception, we also found neurobehavioral correlations between phonetic variables and the amplitude of both early and late ERP components.

Primarly, we found that between-session changes in production correlated with changes in both early (126–146ms) and late (200–220ms) auditory components, though only for specific stimulus steps. We also found a possible relationship between average F1 production in the training session and P2 amplitude in the subsequent identification

task. Thus, the results of the exploratory analyses suggest that that sensorimotor experience may affect early vowel decoding processes (Obleser and Eisner, 2009) as well as later perceptual decision processes (Bidelman et al., 2013; Mostert et al., 2015). The observed pattern of results can be best characterized as a remapping of the relationship between acoustics and articulation by sensorimotor experience. Before arguing for this interpretation, we first consider possible reasons why our group-level results differed from those found previously.

Unlike in Lametti et al. (2014b), our results did not reveal any significant between-group differences in phonetic categorization. Specifically, adapters did not differ significantly from controls with respect to changes in phonetic categorization following speech production training. Furthermore, though the correlation analyses suggested a possible relationship between changes in F1 and changes in phonetic categorization, the direction of the correlation conflicts with that found in Lametti et al. (2014b). However, given that our correlations did not reach significance after applying Holm-Bonferroni corrections, and that they only held when both adapters and controls were included, it is difficult to draw conclusions about the the differences between these two studies.

The results of our correlational analyses suggested that, within each session, F2 was a stronger predictor of overall phonetic categorization than F1 (Fig. 3.3). This may have been due to the fact that the experiment involved Dutch rather than English speaking participants. In addition to a distinction between [ɛ] and [ɪ], Dutch also has a rounded vowel [ʏ], distinguished from [ɪ] by primarily F2 and F3 (Adank et al., 2004). The presence of this phonological contrast may have increased Dutch participant's attention to the value of F2 compared to English participants. It is possible that this attention to F2 may have counteracted any group-level differences induced by changes in F1.

Another confound may have come from the design. Lametti et al. (2014b) utilized a ten-step phonetic continuum, whereas the present study utilized a five-step continuum based on Bidelman et al. (2013). While standardized F1 was found to correlate with changes in phonetic categorization, correlations between this formant and changes in the amplitude of the P2 component were only found for stimulus step one. As a continuum endpoint, this stimulus exhibited almost no differences in behavioral responses before and after training. It may be that clearer between-group differences would have emerged had we utilized a ten-step continuum, increasing the number of stimuli closer to [ɛ].

Finally, both sensorimotor adaptation (Rochet-Capellan and Ostry, 2011; Rochet-Capellan et al., 2012) and phonetic recalibration (Eisner and McQueen, 2005; Reinisch et al., 2014) have been found to be extremely specific. Given that we used a full word ("pet")

during the speaking task, yet presented participants with isolated vowels (as in Bidelman et al., 2013), the lack of group-level effects may be due to differences between the training and test stimuli. It is possible that the effects would have been stronger had we used a "pet-pit" continuum for the perceptual task, or modulated the acoustics of the stimuli to match the gender of the participant.

Returning to the results of the exploratory analyses, changes in the amplitude of the N1b and P2 components during phonetic categorization directly followed speech production training, suggesting a causal relationship between sensorimotor experience and perceptual changes. For the N1b component, this relationship only held for stimulus step three (the most ambiguous stimulus), while for the P2 component, this relationship only held for stimulus step one (clear [ɛ]). While the correlations suggested that this was indeed driven by behavior in the preceding speech production task, it is important to consider whether the results might also be explained by other factors.

One possibility is that the effects reflect exposure to the categorization stimuli. In previous auditory training experiments, simple exposure to a phonetic continuum has been found to elicit increases in P2 amplitude (Tremblay et al., 2010, 2014), which has been associated with the formation of a new phonetic contrast. However, the current study utilized an existing phonetic contrast, and furthermore, found diverging results for participants based on their behavior during the speech production task. When listening to the same [ɛ]-vowel after speech production training, P2 amplitude increased for participants who had produced this vowel with a lower F1 during training, yet decreased for participants who had produced this vowel with a higher F1. Therefore, it is unlikely that the observed effects can be simply attributed to exposure to the phonetic continuum.

Another possibility is that the effects reflect selective adaptation and/or phonetic recalibration in response to the categorization stimuli. Adapters received altered auditory feedback, leading to a mismatch between production and perception as well as a change in auditory feedback, while controls received unaltered feedback. Therefore, it could be argued that changes in amplitude observed in the control participants may reflect selective adaptation, having repeated the same vowel multiple times, while changes in amplitude observed in adapters may reflect phonetic recalibration of the trained vowel (Kleinschmidt and Jaeger, 2015a; Shiller et al., 2009; van Linden et al., 2007). Repeated presentation of the same stimulus (or stimulus type) elicits ERP components with diminished amplitude compared to a novel stimulus (Belin and Zatorre, 2003; Tian and Poeppel, 2010, 2013). While this has previously been ascribed to "fatiguing" of feature detectors (Eimas and Corbit, 1973; Samuel, 1986), recent modeling suggests that speech sound categories may be likened to probability density functions, which are updated on the basis of the distribution of input exemplars (Kleinschmidt and Jaeger,

2015b). Therefore, in controls, repeated production of unaltered [ɛ] may have sharpened feature representation for this vowel, leading to less activity when these features are matched (stimulus one) and increased activity for more ambiguous stimulus steps (e.g., step three, Fig. 3.4D). Conversely, in adapters, the error between expected and heard feedback may have led to a shift in the center of this distribution, leading to increased error during perception. If this were true, then we would expect to observe increases in the amplitude of all components as a consequence of adaptation to altered feedback, and a decrease in the amplitude of these components as a consequence of unaltered feedback. Accordingly, most adapters exhibit increases in P2 amplitude in response to stimulus one. However, most adapters exhibited *decreases* in N1b amplitude in response to stimulus three, while many controls exhibited *increases* in the amplitude of this component (Fig. 3.5D). Furthermore, in a similar experiment, Ito et al. (2016) found decreases in P2 amplitude after adaptation to altered feedback. The opposing effects observed in these results cannot be accounted for by selective adaptation and phonetic recalibration alone.

We therefore propose that changes in component amplitude in both groups reflected changes in the mapping between acoustic values and articulatory features (Poeppel et al., 2008; Tourville et al., 2013). For example, the amplitude of the P2 component has been found to be greater when an ambiguous stimulus is categorized as a low vowel [a] than when this same acoustic stimulus is categorized as a high vowel [u] (Bidelman et al., 2013). As stated in the introduction, if the vowel [ɛ] in "pet" comes to be associated with a more closed or open jaw position due to changes in production (regardless of feedback), then a sensory-to-motor mapping account would predict that P2 amplitude should increase or decrease correspondingly.

In Ito et al. (2016), the value of F1 was decreased for participants in the altered feedback condition, leading adapters to produce the target [ɛ]-vowel with a greater F1 frequency compared to baseline (i.e., as a more [æ]-like vowel, in which the jaw is lower). Greater adaptation responses corresponded to greater *decreases* in the amplitude of the P2 component when listening to the trained vowel. In the present experiment, the experimental manipulation consisted of an increase in the value of F1, leading adapters to produce a more [ɪ]-like vowel. Greater adaptation responses were found to correspond to greater *increases* in P2 amplitude. These correlations suggest that the direction of change in the amplitude of the P2 component during perception of the trained vowel corresponded to those expected when perceiving a vowel with a specific height (Bidelman et al., 2013; Shestakova et al., 2004).

Based on the results of Bidelman et al. (2013), we can re-characterize the observed changes in P2 amplitude in these experiments. Articulating [ɛ] as a higher, more [ɪ]-like

vowel during production (as in the present experiment) caused a previously presented auditory stimulus to be perceived as if it were a lower vowel. Articulating [ɛ] as a lower, more [æ-like] vowel (Ito et al., 2016) led the same vowel to be perceived as if it were a more closed vowel. One might object that this is a result not of the articulatory behavior, but of the auditory feedback participants experienced during training. That is, rather than having a sensorimotoric cause, the results could simply reflect exposure to auditory feedback during production.

But a purely sensory explanation is unable to account for the modulation of the N1b and P2 components observed in adapters. Speakers have been found, in response to similar feedback shifts, to only partially compensate for shifts in auditory feedback (Katseff et al., 2012). In the present study, adaptation opposing the direction of the shifted feedback compensated on average for 38% shift in auditory feedback. This means that despite producing a vowel with a lower F1, participants nevertheless heard themselves producing a vowel with a higher F1 value than normal. Thus, a purely sensory account would predict the opposite direction of correlation compared to that observed. That being said, the variation observed in amplitude of these components was not completely accounted for by the changes in production. Speakers have been found to differ with regard to their dependence on somatosensory compared to auditory feedback in a production task (Lametti et al., 2012). It may be the case that individual differences in sensory preference may have modulated the observed correlation.

Throughout this paper, we have consistently referred to our perceptual task as phonetic categorization rather than speech perception. This deliberate choice reflects the fact that effects observed in phonetic categorization tasks do not necessarily coincide with those observed in other, possibly more natural speech contexts (Hickok et al., 2009; Hickok and Poeppel, 2000, 2004; Krieger-Redwood et al., 2013). Therefore, in order to generalize our results beyond phonetic categorization, it is important to consider how this task may resemble natural speech perception. The current experiment employed a two-alternative forced choice task, leading to a categorical response profile with a rather sharp identification boundary (Bidelman et al., 2013; Chang et al., 2010; Goldstone and Hendrickson, 2010; Liberman et al., 1957). The experimental design therefore led participants to focus on acoustic cues relevant for distinguishing the target contrast. However, the 'categoricalness' of categorical perception may diminish or even disappear when more response options are available (Gerrits and Schouten, 2004; Lotto, 2000; Schouten et al., 2003). This suggests that mapping acoustics onto articulatory representations based on one's own sensorimotor experience may not apply to situations where the range of possible categories to which an acoustic signal can be assigned is more open, as in natural speech. In such cases, it may be better to rely on lexical information to reinterpret acoustics (e.g., Ganong, 1980; Norris et al., 2003). Yet

when context constrains the range of possible sound categories, simulation of candidate phonetic categories may aid perception (Poeppel et al., 2008; Tian and Poeppel, 2013).

We have argued that the cortical and perceptual effects observed in this study reflect a mapping of acoustics onto articulation in order to classify a speech sound, and that these mappings may be updated by recent sensorimotor experience. Therefore, in addition to supporting and maintaining speech production abilities (Lane and Webster, 1991; Niziolek et al., 2013), our results suggest that sensorimotor experience may play a role in certain perceptual contexts as well. As has been noted, "the task of perceiving speech sounds is complex and the ease with which humans acquire, produce and perceive these sounds is remarkable" (Carbonell and Lotto, 2014). It is equally remarkable that humans are able to swiftly and flexibly take advantage of diverse cues and resources in order to deal with the perceptual task at hand (Brown and Kuperberg, 2015; Erb et al., 2013). Listeners have been argued to draw on their knowledge about how speech is produced in order to help decode speech under difficult listening conditions (Nuttall et al., 2016) and as a tool to predict how upcoming speech will sound (Brunellière et al., 2009; Tian and Poeppel, 2010, 2013). What this exploratory study has suggested is that this knowledge is not static, but is updated and modulated by our ongoing sensorimotor experiences.

Chapter 4

Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker

In different tasks involving action perception, performance has been found to be facilitated when the presented stimuli were produced by the participants themselves rather than by another participant. These results suggest that the same mental representations are accessed during both production and perception. However, with regard to spoken word perception, evidence also suggests that listeners' representations for speech reflect the input from their surrounding linguistic community, rather than their own idiosyncratic productions. Furthermore, speech perception is heavily influenced by indexical cues that may lead listeners to frame their interpretations of incoming speech signals with regard to speaker identity. In order to determine whether word recognition evinces similar self-advantages as found in action perception, it was necessary to eliminate indexical cues from the speech signal. We therefore asked participants to identify noise-vocoded versions of Dutch words that were based on either their own recordings or those of a statistically average speaker. The majority of participants were more accurate for the average speaker than for themselves, even after taking into account differences in intelligibility. These results suggest that the speech representations accessed during perception of noise-vocoded speech are more reflective of the input of the speech community, and hence that speech perception is not necessarily based on representations of one's own speech.

This chapter is adapted from:

Schuerman, W. L., Meyer, A., & McQueen, J. M. (2015). Do We Perceive Others Better than Ourselves? A Perceptual Benefit for Noise-Vocoded Speech Produced by an Average Speaker. *PLoS ONE, 10*(7), e0129731.

4.1 Introduction

> Speech production does not operate in a vacuum, free from the influences of its perceptual counterpart; the two processes are coupled and closely linked. (Casserly and Pisoni, 2010)

To someone unfamiliar with the field of speech research, it may seem odd that such a statement needs to be expressed explicitly considering how natural the link between the production and perception of speech appears to be. Given the many dissociations found between speech production and perception, however, it may seem quite reasonable to treat production and perception as separate objects of study. For example, lesions to particular areas of the brain appear to solely hinder production abilities while leaving perceptual abilities intact (and vice-versa (Scott et al., 2013)), and it has been well-documented that children are able to perceive certain sounds that they are not yet able to produce (Brown and Berko, 1960). Theories have indeed been developed for either production alone (Dell, 1986) or perception alone (McClelland and Elman, 1986). But at the opposite extreme, others propose that speech perception crucially relies on production mechanisms (Pickering and Garrod, 2007), in concordance with certain theories of action perception (Prinz, 1990). In this study, we test listeners' perception of a type of spectrally manipulated speech, noise-vocoded speech, in which indexical cues have been largely eliminated. Testing perception of such speech allows us to investigate whether listeners are better at perceiving their own speech or the speech of another person in the absence of primary phonetic cues to speaker identity, and how this may enrich our understanding of the coupling between production and perception mechanisms.

Models of speech perception that argue that the production system plays a critical role, such as the motor theory of speech perception (Liberman et al., 1967) and analysis by synthesis (Halle and Stevens, 1959), have existed for some time and have undergone many revisions and changes (Galantucci et al., 2006). Such theories have recently experienced a resurgence in interest due in great part to the growing body of neurobiological evidence demonstrating that speech motor areas (ostensibly only supposed to be activated during production) may become activated not only during production but also during perception (Poeppel and Monahan, 2011). Much of this neurobiological evidence has been gathered from experiments involving transcranial magnetic stimulation (TMS), in which targeted magnetic pulses are used to stimulate or inhibit activity in select areas of the brain (Devlin and Watkins, 2007). By pairing TMS pulses with electromyography, which measures the activation of muscles via electrodes placed on the surface of the skin, it becomes possible to measure the degree of motor activation during the production or perception of an action. For example, muscle activation in lip

regions involved in articulation has been found to be greater when listening to speech vs. non-speech (though only in response to left-hemisphere stimulation) (Watkins et al., 2003). More recent studies suggest that motor involvement may not always be necessary in order to perceive speech, but instead that motor areas may be recruited only during the perception of noisy, ambiguous, or non-native speech (Scott et al., 2009). However, as proponents of motor involvement in speech perception have argued, pristine speech such as that found in laboratory conditions may be more likely to be the exception rather than the norm (Bartoli et al., 2015).

While neurobiological evidence suggests that motor areas may become activated during speech perception, it remains unclear what the functional role of this activation may be. One possibility is that this activation indicates parity between the mental representations accessed during production and perception. According to this view, perceiving an action requires accessing the same mental representations utilized when producing that action. This is in effect an argument for the common coding theory of perception, which argues that "percept codes and act codes are formed in the same format and are stored or maintained in a common representational medium" (Prinz, 1990, p. 171). One of the predictions of this theory is that perception is facilitated according to the degree to which the incoming sensory input matches the internal "action" representation; the more the action that we are perceiving matches with how we ourselves would produce it, the more easily the corresponding representation is activated.

Evidence in support of common coding has been provided by studies that have found perceptual facilitation for self-generated actions. For example, in a study by Knoblich et al. (2002), participants were asked to write either the number "2", or the first stroke in the number "2" without the horizontal line (a hook), utilizing a stylus and tablet. When later presented with recordings of only the hook stroke and asked to predict whether a horizontal stroke would follow (i.e. whether this was an instance of drawing the number "2" or simply drawing the hook part), participants were found to perform above chance accuracy when the stimulus was drawn from their own productions but only at chance when it had been produced by another participant. Similar perceptual advantages for self-generated stimuli have also been found for predicting where a thrown dart will land (Knoblich and Flach, 2001). The authors argue that these results support a common coding account for action representations; our perception of some event, and the subsequent predictions that follow, are facilitated according to the degree of similarity between the actions that we observe and our own production of those actions.

Few studies have been undertaken to determine to what extent these effects may hold true in the realm of speech perception. If representations for speech sounds are shared for both production and perception, this predicts that speech perception ought to be

facilitated for self-generated speech as well. A recent experiment on the perception of lip-read speech suggests that this may indeed be the case (Tye-Murray et al., 2013). Two groups of ten participants were videotaped reading approximately 360 sentences aloud. These same participants were later tested on a subset of the 360 sentences, with an equal number of video clips drawn from every participant in the group (including each participant's own recordings). Accuracy was scored as the percentage of words correctly identified. Fifteen of the 20 participants correctly identified a higher percentage of words when lip-reading themselves than when lip-reading other participants. This is rather startling, as speakers do not usually observe their own articulations and receive only auditory, tactile and proprioceptive feedback online. The results suggest that listeners do in fact utilize their production experience during perception, yet it is unclear to what extent this is restricted to the task of lip-reading. If auditory speech perception were also a form of action perception, this would predict that similar self-advantages ought to be found in auditory word recognition as well.

However, when comparing action perception to speech perception, it is crucial to recognize that speech perception is known to be influenced by indexical cues to speaker identity. For example, the perception of synthesized fricatives on a continuum between "sod" and "shod" can be influenced by the perceived gender of the speaker (Strand and Johnson, 1996). In a subsequent experiment, in which listeners were presented with audio stimuli that had been generated from the recordings of a speaker whose voice was judged to be non-prototypical for their gender were paired with video of either a male or female speaker, it was found that the visual cue to speaker gender further modulated the perceived fricative boundary. This demonstrates that auditory cues to gender and audio-visual integration of perceived gender can influence the perception of speech sounds. Later experiments replicated this audio-visual gender integration effect for vowel stimuli and further found that with audio-only presentation, the phoneme boundary could be modulated simply by instructing the listener to imagine the speaker as either male or female (Johnson et al., 1999). Shifts in phoneme categorization have also been found when participants associate a speaker with a specific social group (e.g. nationality) on the basis of an accompanying label (Niedzielski, 1999). These studies demonstrate that speech perception can be influenced by the perceived identity of the speaker as indexed by both linguistic and extra-linguistic cues. It is therefore not unreasonable to assume that recognizing stimuli as self-produced may also influence perceptual processing. This means that any benefit in recognizing one's own speech could be the result not only of overlap in representations, as predicted by the common coding account, but also because the listener's own speech was recognized as such.

If, in contrast, representations accessed for perception are fundamentally distinct from

those accessed for production, those representations may be more reflective of the statistical properties of the speech in a given linguistic community. It is evident from experiments on cross-linguistic speech perception that perceptual identification of a speech sound depends on the sound's distribution in the perceiver's own language. For example, when tested on synthesized speech samples along a continuum between /l/ and /r/, American English speakers show categorical discrimination effects which reflect the bimodal distribution of these speech sounds in American English, whereas inexperienced Japanese speakers do not show a categorical discrimination effect (MacKain et al., 1981). Experienced Japanese listeners, however, who have been exposed to tokens of /l/ and /r/, and thus a bimodal distribution of sounds, perform similarly to the native English speakers. Studies differ with regard to whether perceptual effects are coupled to production ability; Bradlow et al. (1999) found that Japanese speakers given intense perceptual training on the English /l/ $\sim$ /r/ distinction improved both their perception and production of the contrast, even up to three months following training. This suggests that perceptual skills may be tightly coupled to production skills. In contrast, Sheldon and Strange (1982) found that participants were often better at producing a given phoneme (as judged by native English speakers) than they were at perceiving the same phoneme produced by themselves or other Japanese learners. At first glance, this would seem to suggest dissociation between production and perception. However, this study also found that four of the five Japanese participants made fewer errors when identifying /l/ and /r/ recordings produced by themselves compared to stimuli produced by other Japanese speakers or native English speakers. Further research by Borden et al. (1983) on Korean speakers suggests that individuals may differ with respect to whether perception is more accurate than production or vice-versa, though this may be related to length of time spent in an English-speaking environment (Sheldon, 1985). These cross-linguistic perception studies thus provide conflicting evidence with regard to the common coding of speech representations.

At a broad level of granularity, speakers' productions will almost necessarily reflect the statistical properties of their native language. It is therefore difficult to determine from cross-linguistic experiments whether decoding incoming speech in one's native language utilizes representations that are more reflective of the listener's own production idiosyncrasies or of the statistical properties of the speech community, as such fine-grained differences will most likely reside at the sub-phonemic level. In order to investigate which of these two alternatives may be the case, it is necessary to examine within a single language whether listeners show enhanced recognition when listening to themselves or to a "typical" speaker whose productions more closely approximate the average of their linguistic community.

4.2 Experiment 1

In this study, we investigated whether spoken word recognition is facilitated more when words have been generated by the participants themselves or more when they were generated by a statistically average speaker. Experience listening to a single talker leads to an increase in intelligibility for subsequent stimuli from the same talker (Bradlow and Bent, 2008), even when this signal is extremely altered. For example, Remez et al. (2011) examined the recognition of isolated sine-wave sentences. Accuracy was found to be greater when the test sentences and exposure sentences were produced by the same talker in the same modality rather than by different talkers. This suggests that even when the speech signal is distorted, listeners are still sensitive to talker-specific phonetic regularities and that exposure to these regularities can facilitate recognition of novel words produced by the same talker.

Due to the extensive experience speakers have with their own productions, we might expect speakers to be very sensitive to their own phonetic idiosyncrasies. This seems likely given that speakers have been shown to correct online for subtle (sub-phonemic) deviations from their intended speech targets (Niziolek et al., 2013). Listening to one's own voice via a recording is, however, a very different experience from monitoring one's productions. When listening to recordings of our own voices, we received a signal only via air-conduction. This contributes to the experience of hearing our voices as 'higher' on a recording compared to how we hear ourselves during speech production. During active speech production, a speaker hears their own voice via both air- and bone-conduction, resulting in a different psychoacoustic experience compared to when a speaker listens to recordings of their own voice (Békésy, 1949). Thus, the spectral properties of self-produced speech are shifted when heard on a recording compared to what is heard during production.

However, certain phonetic properties, such as the length of a given stop burst or changes in amplitude, remain invariant between listening to one's own voice while speaking and listening to recordings of one's own voice. We therefore decided to utilize "noise-vocoded speech" (NVS) to examine the perception of self-produced speech. NVS preserves temporal and amplitude cues while eliminating fine-grained spectral detail (Shannon et al., 1995). This includes many of the primary cues, such as formant dispersion and pitch, that people use to distinguish voices from one another (López et al., 2013). Many recent studies utilize NVS to manipulate the intelligibility of linguistic stimuli systematically. These studies have found that, even with relatively high levels of degradation, participants are quickly able to adapt to the spectral manipulation and accurately recognize the original utterance (Davis et al., 2005; Hannemann et al., 2007; Sohoglu

et al., 2012). Importantly, several studies have found that talker-identification in NVS is greatly impaired (Büchner et al., 2009; Turner et al., 2004; Vongphoe and Zeng, 2005; Zhang et al., 2010). This rather extreme manipulation therefore provides another benefit, in that it eliminates many linguistic cues to speaker identity that influence speech perception (Johnson et al., 1999; Strand and Johnson, 1996). It thus allowed us to ask whether there was a self-advantage in speech perception independent of listeners' ability to recognize the speech as their own.

In Experiment 1, we examined listener accuracy for identifying single NVS words that had been produced either by the participants themselves or by a statistically average speaker. The experiment consisted of two sessions. First, in the recording session, native Dutch participants produced 120 Dutch words. After comparing the phonetic properties of the recordings, we selected from amongst the participants one speaker who differed least from all other participants in the sample. This model speaker acted as a proxy for a statistically average speaker of the linguistic community. The recordings were then converted into NVS stimuli. After a one-week interval, the same participants were presented with these NVS stimuli and asked to identify the original words. Participants were unaware that half of the stimuli were based on their own recordings and that the other half were based on the recordings of the selected average speaker.

If identical representations for spoken words are utilized for both production and perception, as common coding posits, then we would expect to find greater accuracy for self-produced NVS stimuli. Furthermore, if listeners' access to these representations does not depend on the speaker's identity, we would expect to find advantages for self-produced stimuli regardless of perceived speaker identity. However, if representations are more representative of the overall input of the linguistic community, then we predict that participants would be more accurate at identifying words generated by a speaker whose productions more closely align with this average.

4.2.1 Materials and Methods

4.2.1.1 Participants

Twenty-eight female native speakers of Dutch between the ages of 19 and 26 (μ = 21.63), all with healthy vision and hearing, participated in the experiment. In order to minimize large between-speaker differences, only female participants were invited to participate. All 28 participated in the production task. One participant was selected to be the "average" model speaker, leaving 27 participants for the identification task, which was performed approximately one week after the production task.

4.2.1.2 Ethics Declaration

Ethical approval for this study was obtained from the Ethics Committee of the Social Sciences Faculty of Radboud University. Written consent was obtained from each participant on the first day of the study. Participants were informed that their participation was voluntary and that they were free to withdraw from the study at any time without any negative repercussions and without needing to specify any reason for withdrawal. All were reimbursed for their participation.

4.2.1.3 Stimuli

The stimulus set of 120 Dutch words was based on English materials developed for speech-in-noise recognition experiments by Bradlow and Pisoni (1999). 35 phonemic segments were represented in this set of words, though with different frequencies (e.g., /z/ appears in 10 words, while /b/ appears in only 6 words). Words in the stimulus set were divided into two groups according to frequency, phonological neighborhood density, and the average frequency of their phonological neighbors. Phonological neighbors were defined as words that differed from the target word by the deletion, addition, or substitution of a single phoneme. "Easy" words were those that had high frequency (μ = 226.53 per million), few phonological neighbors (μ = 9.2), and low neighborhood frequency (average frequency of all phonological neighbors; μ = 46.49 per million). "Hard" words, on the other hand, had relatively low frequency (μ = 16.71), larger phonological neighborhoods (μ = 20.85), and higher neighborhood frequency (μ = 470.98 per million). Thus, Easy words were likely to stand out as there were no or relatively few similar sounding words, while Hard words were more likely to be confused with more frequent, similar sounding words. Pilot tests confirmed that participants correctly recognized noise-vocoded versions of the Easy words more often than Hard words, validating the use of these parameters to distinguish the two conditions. All words utilized in this experiment were classified as 100% familiar according to a 559-participant age-of-acquisition study (Ghyselinck et al., 2000).

4.2.1.4 Production Task

During the recording session, participants were comfortably seated in a sound-attenuated booth. A pop-filter shielded microphone was placed approximately 10 cm away from each participant's mouth. Participants were presented visually with the 120 stimulus words on a computer screen and asked to read them out loud in a "normal manner"

within a specific recording window that was signaled visually. A black screen was presented for 500 ms, after which the target word appeared on the screen above a white box. After 250 ms a red circle appeared within the white box to mark the beginning of the recording and signal the participant to read the word out loud. Recording continued for 1750 ms, after which the red circle disappeared. At the end of each such trial, the participant was given the option to repeat the trial (e.g. if the participant had coughed during the recording) or to continue on to the next trial. The order of stimulus presentation was fully randomized for each participant. All productions were recorded using Presentation software (Version 0.70, www.neurobs.com).

4.2.1.5 Model Speaker Selection

Each recorded sound file was segmented and analyzed utilizing Praat (Boersma and Weenink, 2013). Measurements taken from the participant recordings consisted of average word duration per speaker, average segment duration per speaker, duration of each word (120 variables), segment duration by word (e.g. /z/ in 'zuil'; 426 variables), and average segment amplitude (averaged across words; 35 variables). In addition, segments were also analyzed with respect to the NVS manipulation, which divided the sound file into six frequency bands. Therefore we included measurements of average amplitude by segment by frequency band (e.g., the average amplitude of each frequency band for all words in which the given segment appears; 210 variables) and standard deviation by segment by frequency band (210 variables). In order to find the most "average" speaker, participants were compared according to the aforementioned variables. Due to the fact that many of these variables were highly correlated, standard principal component analysis was used to reduce the number of dimensions for comparison from 1003 to 27. Euclidean distances between participants were calculated based on these 27 components, and the participant with the smallest average distance from all other participants was chosen as the average speaker.

4.2.1.6 Stimulus Preparation

The intensity of each of the total 3360 sound files (120 words * 28 participants) was normalized by root-mean-squared amplitude. The normalized sound files were then transformed into 6-band noise-vocoded speech (Shannon et al., 1995). These six spectral frequency bands corresponded to equally spaced distances along the basilar membrane (Greenwood, 1990). Pre-experiment pilots determined that this level of degradation avoided ceiling and floor performance in the identification task.

4.2.1.7 Identification Task: Design

For each participant, each of the 120 target words was assigned to only one of two "Talker" conditions, a "Self" condition or an "Average Speaker" condition. The number of words from the Easy and Hard word lists was balanced across these Talker conditions (30 each). This ensured that each Talker condition included a range of words with equivalent and variable difficulty. Repetition of words across Talker conditions was avoided due to the possible confounding effect of acclimatization to the noise-vocoding manipulation. Order of stimulus presentation was randomized within the two Talker blocks. Order of block presentation was counterbalanced across participants.

4.2.1.8 Identification Task: Procedure

In the identification phase, participants attempted to identify noise-vocoded versions of the recorded words. A trial began with visual presentation of a fixation cross located in the center of a computer screen. The auditory stimulus was presented after 250 ms of silence. Following stimulus presentation, participants attempted to identify the target word by typing in their response via the computer keyboard. Once the participant confirmed their response by pressing the Enter key, the next trial began.

A 10-word practice session preceded the experimental blocks in order to familiarize participants with the noise-vocoded stimuli and the task. Feedback was given after each practice item, indicating whether or not the participant had correctly guessed the word. None of the words in the practice session appeared in the experimental blocks. Participants received no feedback during test blocks.

4.2.2 Results

Participants responded using standard Dutch orthography; given that Dutch orthography may represent the same sound in multiple ways (e.g., word final [t] can be spelled as either "t" or "d"), target words and participant responses were transcribed into a standardized broad phonemic script. Correspondence between target and response was measured utilizing these phonemic transcriptions rather than the raw orthographic input. The average percentages of correct responses (defined as 100% match between target and response transcription) by Talker and Word Difficulty are displayed in Fig. 1. While individuals varied with respect to their accuracy, responses were much more accurate for Easy words (μ = 0.57, SE = 0.012) than Hard words (μ = 0.37, SE = 0.012) and moderately more accurate for words in the Average Speaker condition (μ =

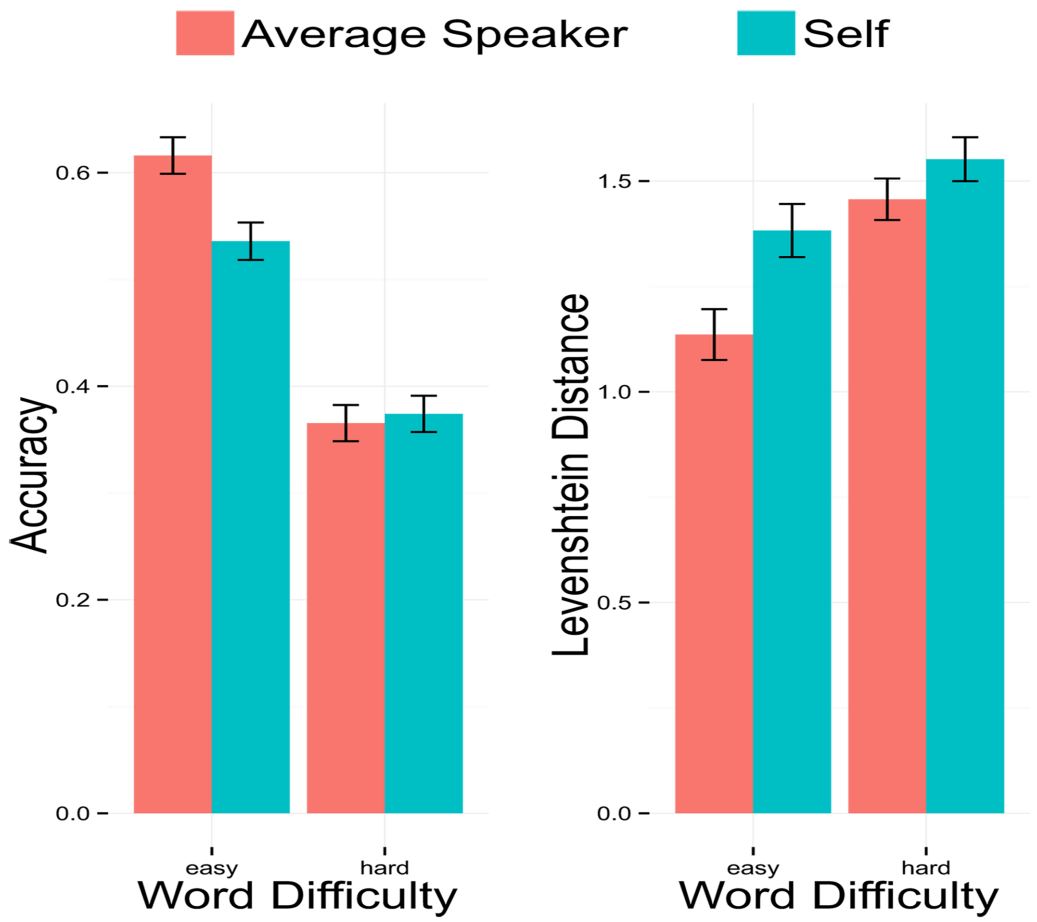


FIGURE 4.1: **Accuracy and Levenshtein Distance results.** Average proportion of correct answers (Accuracy) and Levenshtein Distance for each word list, according to Talker, with standard error bars. For Levenshtein Distance, higher scores indicate greater inaccuracy.

0.49, SE = 0.012) than in the Self condition (μ = 0.45, SE = 0.012). For Easy words, participants were more accurate in the Average Speaker condition (μ = 0.62, SE = 0.017) than in the Self condition (μ = 0.53, SE = 0.017). For Hard words, accuracy was slightly greater in the Self condition (μ = 0.37, SE = 0.017) than in the Average Speaker condition (μ = 0.36, SE = 0.017).

Scoring responses in a binary format as either correct or false depending on a 100% match between target and response transcription fails to capture the detail contained within the transcribed responses. If the target word was “vacht” (fur) and a participant reported to have heard “macht” (power), then the participant did in fact accurately identify a substantial portion of the target word. For this reason, Levenshtein Distances (Levenshtein, 1966) were computed between the target and response transcriptions. Levenshtein Distance is a metric that measures the number of edits (substitutions, additions, deletions) required to transform one string of characters into another. This

method has been utilized, for example, to measure the linguistic distance between dialects of a given language (Beijering et al., 2008). According to this metric, a perfectly correct response is scored as 0 (no edits), while larger scores indicate greater inaccuracy.

Average Levenshtein Distance measurements corresponded to the binary scored accuracy measurements; participants showed greater inaccuracy (higher average Levenshtein Distance) for Hard words ((μ = 1.50, SE = 0.03) than for Easy words (μ = 1.26, SE = 0.04), as well as greater inaccuracy for words in the Self condition (μ = 1.47, SE = 0.04) than words produced by the Average Speaker (μ = 1.30, SE = 0.04). As Fig. 1 shows, for Easy words, average Levenshtein Distance was greater in the Self condition (μ = 1.38, SE = 0.06) than in the Average Speaker condition (μ = 1.16, SE = 0.06); for Hard words, average Levenshtein Distance was also greater in the Self condition (μ = 1.55, SE = 0.05) than in the Average Speaker condition (μ = 1.46, SE = 0.05). An examination of each individual's scores reveals that while some participants showed no difference in inaccuracy between Talker conditions and certain participants were more accurate for Self-produced stimuli, most participants were more accurate for the Average Speaker's sound files than their own (Fig. 2).

Due to the high number of zero scores (100% accurate responses) present in the data, we utilized a hurdle model to analyze the Levenshtein distance findings. A hurdle model is a combination of two models; it combines a binomial regression model to analyze zero vs. non-zero responses, as well as a zero-truncated Poisson model to analyze all responses greater than zero (i.e. those with at least one error). Given that the words varied between Talker conditions and were produced by different speakers, we decided to analyze the results using mixed-effects regression. All analyses were conducted in R using the glmmADMB package (Fournier et al., 2012).

All statistical models included the maximal random effects structure justified by the experimental design (Barr et al., 2013). This consisted of random intercepts for Item and Participant, random slopes for Talker (Self/Average Speaker) by Word, as well as random slopes for Talker, Word Difficulty, and the interaction between Word Difficulty and Talker by Participant.

We report the predictors entered into the binomial (*Binom*) and truncated Poisson (*TruncPoiss*) model at each step. The maximal fixed-effect structure was determined by likelihood comparison. Model comparison began with null models containing only the random effect structure. Next, we added a fixed effect of Word Difficulty type, which was found to substantially improve the fit of both models (*Binom*: AIC = 3767.66, BIC = 3816.33, LogLik = -1875.8, χ^2 = 19.18, $p < 0.001$; *TruncPoiss*: AIC = 5223.34, BIC = 5266.88, LogLik = -2603.7, χ^2 = 14.34, $p < 0.001$). Addition of a fixed effect of Talker

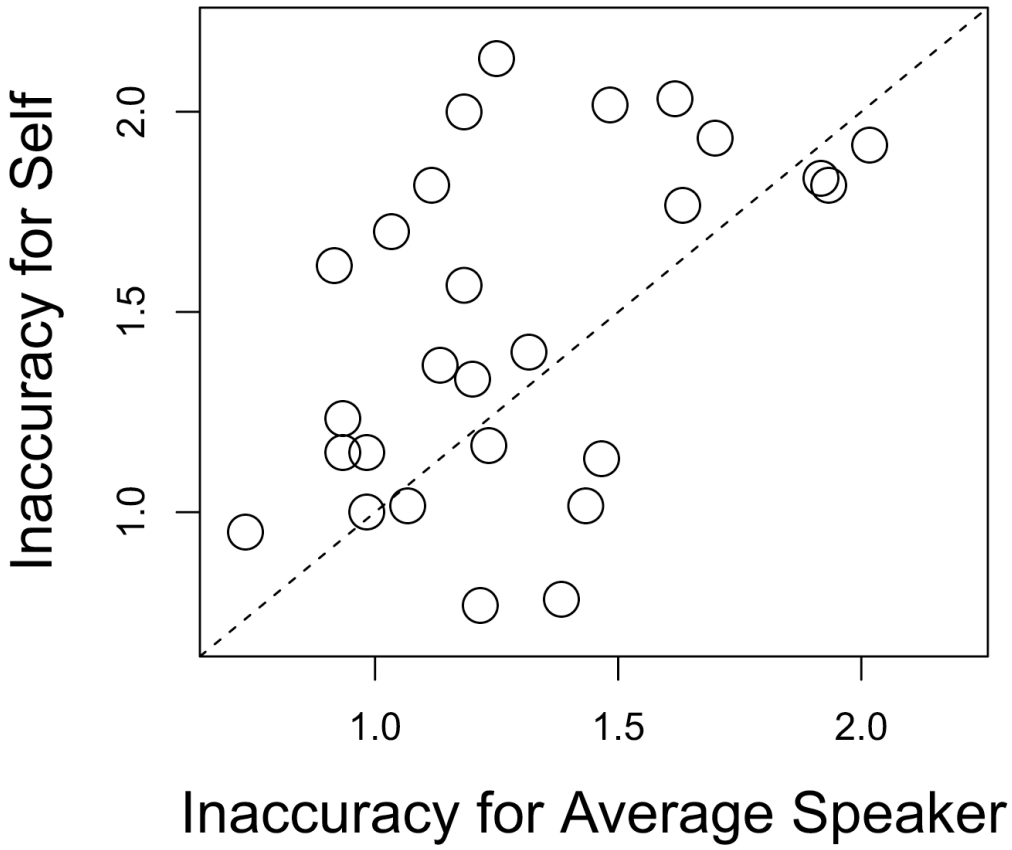


FIGURE 4.2: **Average Levenshtein Distance for stimuli produced by the participant (Self) and by the Average Speaker.** Higher scores indicate greater inaccuracy.

also significantly improved both models (*Binom*: AIC = 3764.62, BIC = 3819.37, LogLik = -1873.3, χ^2 = 5.04, $p < 0.05$; *TruncPoiss*: AIC = 5221.02, BIC = 5270.01, LogLik = -2601.5, χ^2 = 4.32, $p < 0.05$). Inclusion of an interaction term between Talker and Word Difficulty did not significantly improve either model.

Parameters for the fixed effects are reported in Table 1 for the final versions of the binomial and truncated Poisson model, both with reference categories of "Easy" (Word Difficulty) and "Average Speaker" (Talker). For the binomial model, the fixed effect of Word Difficulty reveals that when comparing zero (correct) and non-zero (incorrect) responses, the likelihood of a non-zero response was significantly greater for more difficult words. However, for the truncated Poisson model, the negative slope indicates that Levenshtein Distance decreased in the Hard condition, possibly because of the large number of trials in this condition in which there was only one error between target and response transcription.

TABLE 4.1: **Results of the binomial and truncated poisson mixed-effects regression analyses.**

	Binomial			
Fixed Effects:	Estimate	Std. Error	Z-value	Pr($>\|z\|$)
(Intercept)	-0.530	0.194	-2.73	0.0063**
Word Difficulty (Hard)	1.15	0.253	4.55	<5.3e-06***
Talker (Self)	0.264	0.114	2.32	0.0203*
	Truncated Poisson			
Fixed Effects:	Estimate	Std. Error	Z-value	Pr($>\|z\|$)
(Intercept)	0.874	0.066	13.24	<2e-16***
Word Difficulty (Hard)	-0.312	0.080	-3.79	0.0001***
Talker (Self)	0.075	0.036	2.08	0.0379*

Fixed effects reported in reference to Talker (Average Speaker) and Word Difficulty (Easy).

4.2.3 Discussion

The results suggest that participants were more accurate at identifying noise-vocoded words produced by an average speaker than by themselves. Differences between Talker conditions were greater for Easy words than Hard words. This is somewhat surprising because factors that would improve intelligibility would be expected to have a stronger effect when stimuli are more difficult to decode. For example, in a study on speech recognition in noise, Bradlow and Pisoni (1999) found greater differences between easy and hard words at a fast speech rate than at a slow speech rate. In the same study, however, the authors also compared speakers' perception of easy and hard words in single and multi-talker conditions; they found that the differences in accuracy between easy and hard words was greater in the single-talker condition than in the multi-talker condition. The authors argue that the increase in accuracy in the single talker condition stems from listeners' "ability to take advantage of consistent surface information about a particular talker's voice" [p. 11]. These results mirror our own; we find greater differences between Easy and Hard words in the Average Speaker condition compared to the Self condition. This may suggest that the typicality of the speaker (i.e. how closely the speaker's productions align to the statistical average of the community) may impart an advantage similar to that provided by repeated exposure to a single talker's voice.

Due to the fact that each participant listened to a different set of sound files (Self vs. Average Speaker), differences in intelligibility across talkers may have led to the observed pattern of results. For example, with respect to the observed individual variation (Fig. 2), it may have been the case that the participants who showed greater accuracy for their own sound files than for those produced by the Average Speaker had simply enunciated more clearly during the recording session; conversely, those participants who

were more accurate in the Average Speaker condition than in the Self condition may have enunciated less clearly or their recordings had been rendered more unintelligible by the noise-vocoding procedure. If so, we would expect to find no effect of Talker after having accounted for the general intelligibility of each speaker.

In order to assess this possibility, we conducted a control experiment to determine the average intelligibility of each speaker. Noise-vocoded stimuli from each talker in the main experiment were presented to new participants, who performed the same open-response identification task. The percentage of correctly identified words constituted a given talker's "intelligibility score." We predicted that intelligibility scores from this control experiment would correlate with a participant's average accuracy in the Self-condition in Experiment 1. The main purpose of Experiment 2, however, was to establish whether the advantage for the Average Speaker's words in Experiment 1 would remain after talker intelligibility was taken into account.

4.3 Experiment 2: Determining Speaker Intelligibility

4.3.1 Materials and Methods

4.3.1.1 Participants

Sixteen female native speakers of Dutch, between the ages of 19 and 28 (μ = 22.63), all with reported healthy vision and hearing, took part in this control experiment. To maximize similarity to Experiment 1, only female participants were invited. None of them had participated in Experiment 1. All were paid for taking part in the experiment.

4.3.1.2 Stimuli & Design

A set of 112 noise-vocoded words were selected from the original set of experimental materials. In order to account for the variability in word difficulty, these 112 words were divided into four groups based on average by-word accuracy in Experiment 1 (high accuracy, mid-high accuracy, mid-low accuracy, and low accuracy). Each control participant was presented with four stimuli from each talker (one from each word difficulty group), randomly selected from the recordings obtained from the 28 talkers from Experiment 1. This resulted in a total of 64 words per talker (16 per word difficulty group), and ensured that any word level variations in intelligibility would not affect aggregated intelligibility scores. Stimuli were presented in two blocks, with each block repeated

twice (ABAB). Order of presentation was pseudo-randomized such that no Talker was repeated twice in a row.

4.3.1.3 Procedure

The identification task was the same as in Experiment 1.

4.3.2 Results

In order to determine whether the participants differed greatly with respect to how intelligible they perceived each talker to be (defined by average Levenshtein Distance), t-tests were performed on the cumulative distribution of results after each new participant was added in order to determine whether this distribution changed significantly. For example, the average scores for participants one through three were compared to the scores for participants one through four. After four control participants, the *t*-statistic (df = 52.592) was less than 1, suggesting that already after only a few control participants, incorporating results obtained from a new control participant did not significantly alter the distribution of results. Comparing the distribution of all 16 participants to the distribution of the 15 previous participants resulted in a *t*-statistic of -0.33 (df = 54). The negligible difference between these two distributions allowed us to assume that with 16 participants we had achieved a reasonably reliable measure of the intelligibility of each participant that was unlikely to change by adding additional control participants. Testing was therefore terminated at this point. Average talker intelligibility given all 16 participants' data (as measured by Levenshtein Distance) was 1.6 (sd = 0.35), with a minimum (most intelligible) of 0.78 and a maximum (least intelligible) of 2.16.

Accuracy in Experiment 1 in the Self-Condition is plotted against Intelligibility in Fig. 3. As expected, one-tailed t-tests confirmed a significant correlation between accuracy for self-produced stimuli and the intelligibility ratings obtained in Experiment 2 (Pearson's r = 0.66, $t(25)$ = 4.388). The positive correlation reveals that if the control participants found a given talker's speech to be more intelligible, then that talker was more likely to have recognized their own speech more accurately in Experiment 1.

The main purpose of obtaining intelligibility scores for each participant was to determine whether the effect of Talker found in Experiment 1 could be solely explained by differences in participant intelligibility. Therefore the data from Experiment 1 was reanalyzed with intelligibility scores included in the model. A fixed effect of Intelligibility was found to significantly improve the null binomial and null truncated Poisson models (*Binom*: AIC = 3777.66, BIC = 3826.33, LogLik = -1880.8, χ^2 = 9.18, $p < 0.01$;

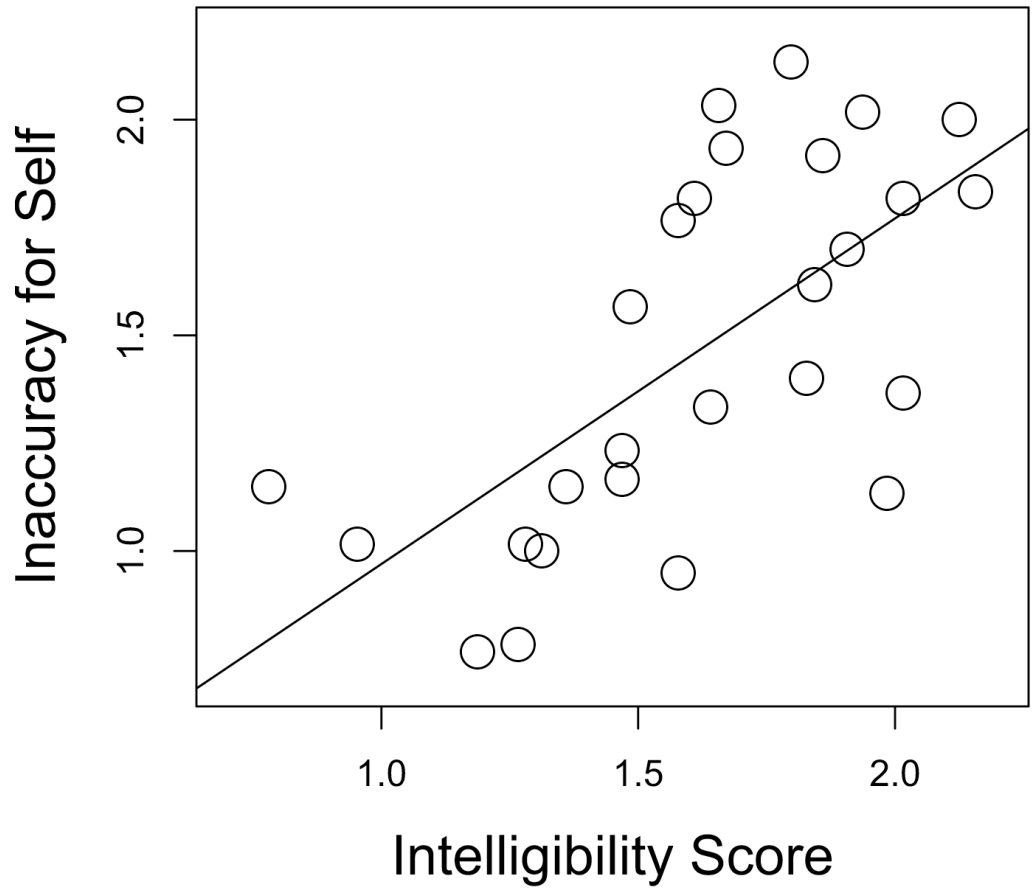


FIGURE 4.3: **Intelligibility ratings.** Intelligibility ratings by Levenshtein Distance for each speaker (Experiment 2) plotted against that speaker's inaccuracy for Self-produced stimuli (Experiment 1), with regression line

TruncPoiss: AIC = 5232.72, BIC = 5276.26, LogLik = -2608.4, χ^2 = 4.96, $p < 0.05$). Crucially, after adding fixed effects of Word Difficulty and Intelligibility, the fixed effect of Talker still significantly improved the fit of both models (*Binom*: AIC = 3756.58, BIC = 3817.4, LogLik = 1868.3, χ^2 = 6.00, $p < 0.05$;*TruncPoiss*: AIC = 5218.42, BIC = 5272.85, LogLik = -2599.2, χ^2 = 4.08, $p < 0.05$). The slope of the fixed effect for Talker remained relatively unchanged in both models (Table 2; *Binom*: β = 0.267, Z = 2.54, $p < 0.05$;*TruncPoiss*: β = 0.07, Z = 2.02, $p < 0.05$) . No significant interaction between Talker and Word Difficulty was found. This analysis reveals that greater accuracy for identifying words produced by the model speaker compared to the participant's own productions, as observed in Experiment 1, cannot be accounted for simply by differences in intelligibility between each participant's noise-vocoded words and those of the model speaker.

TABLE 4.2: **Results of the binomial and truncated poisson mixed-effects regression analyses.**

	Binomial			
Fixed Effects:	Estimate	Std. Error	Z-value	Pr($>\|z\|$)
(Intercept)	-1.722	0.396	-4.35	<1.42e-05**
Intelligibility	0.734	0.213	3.45	0.00057***
Word Difficulty (Hard)	1.156	0.254	4.56	<5.2e-06***
Talker (Self)	0.267	0.105	2.54	0.011*
	Truncated Poisson			
Fixed Effects:	Estimate	Std. Error	Z-value	Pr($>\|z\|$)
(Intercept)	0.579	0.148	3.93	<8.64e-05***
Intelligibility	0.181	0.080	2.25	0.02*
Word Difficulty (Hard)	-0.310	0.080	-3.87	0.00011***
Talker (Self)	0.073	0.036	2.02	0.04*

Intelligibility ratings are based on Levenshtein Distance, therefore greater values indicate that the speaker is less intelligible.

4.4 General Discussion

In this study, we investigated whether the speech representations accessed during perception of noise-vocoded speech are more reflective of an individual's own productions or of the statistical average of the individual's speech community. Common coding theory (Prinz, 1990) predicts that perception should be more accurate when the listener is presented with self-generated stimuli, as has been found in experiments on the perception of self vs. other-generated actions (Knoblich and Flach, 2001; Knoblich et al., 2002; Tye-Murray et al., 2013). If, in contrast, speech representations accessed during perception of noise-vocoded speech are distinct from those used in production and reflect exposure to a range of speakers, listeners should be more accurate at identifying stimuli generated by a speaker whose productions more closely approximate the statistical average of their community. We therefore tested participants on the recognition of noise-vocoded words that had either been produced by themselves in an earlier recording session or by a "statistically average" speaker chosen from the participant population. The pattern of results suggests that most speakers were more accurate at identifying words produced by the statistically average speaker than their own recordings. The results of the control experiment suggest that this advantage for the statistically average speaker was still significant even after taking into account intelligibility differences between the participants' noise-vocoded words.

Given that self-advantages have been found in the visual modality, especially for lip-reading (Tye-Murray et al., 2013), why is it that we do not observe these effects in auditory word recognition? Noise-vocoding was utilized as a means of systematically

degrading the stimuli while preserving temporal and intensity-related idiosyncrasies. It may have been the case that this manipulation eliminated crucial phonetic cues that listeners make use of when both monitoring their own speech and when listening to others. However, this seems unlikely, since the noise-vocoding manipulation preserved talker-specific differences in temporal parameters that listeners are known to be sensitive to, such as voice-onset time (Allen and Miller, 2004). Furthermore, if listeners were not sensitive to variations in these parameters, we would expect to find no difference in accuracy between the Talker conditions independent of general intelligibility.

In experiments on action perception using visual stimuli, participants were often aware of the identity of the producer (Knoblich and Flach, 2001; Tye-Murray et al., 2013). Therefore, it may be the case that self-advantages are dependent on the perceiver consciously recognizing the speech as self-produced. In addition to degrading the stimuli in order to increase the difficulty of word recognition, noise-vocoding was used in our study as a means of preventing possible influences on perception arising from perceived speaker identity (Johnson et al., 1999; Niedzielski, 1999; Strand and Johnson, 1996); noise-vocoding eliminates phonetic cues often used to identify speakers (López et al., 2013). This appears to have been effective, because during debriefing only two participants guessed that they had at some time heard their own voices. When asked what led them to this realization, both participants reported that they had remembered pronouncing a certain word with precisely the same duration during the recording session. Neither of these participants, however, recognized that the study had presented one talker per block, and the majority of the participants did not notice that voices differed between blocks. In the experiments conducted by Knoblich et al. (2002), in which participants attempted to predict aspects of handwriting strokes produced by themselves and other participants, advantages for self-produced stimuli were found without explicit cues to author identity. Knoblich and Flach (2001) argue that self-advantages should not depend on the perceiver's awareness of whether the stimuli were self-produced. Furthermore, they argue that when information about the source of an observed action is sparse, imagined actions are more easily integrated, causing self-other differences to emerge even more swiftly than when identity cues are present. Fundamentally, the common coding theory (Prinz, 1990) predicts that self-advantages should not depend on awareness of the identity of the producer. If participants automatically compare incoming speech to their own production representations, we would expect to find an advantage for self-produced speech regardless of whether the listener is able to identify the voice. Our results suggest, in contrast, that in the absence of cues to speaker identity, listeners decode incoming productions with reference to the statistical average of their community.

The use of NVS stimuli was also likely to have eliminated any phonetic cues that listeners unconsciously use to frame incoming speech with respect to speaker identity. Therefore it remains possible that, were such cues present in the stimuli, listeners would have performed differently. For example, Remez et al. (2011) found that when listeners were given exposure to sine-wave sentences, they performed better when later attempting to identify single sine-wave words produced by the same talker. However, participants who were given exposure to clear versions of the same sentences produced by the same speaker did not show any benefit of exposure; both speaker-specific and modality-specific experience were necessary in order to facilitate word recognition. Nevertheless, as mentioned above, common-coding suggests that the same representations are accessed during both production and perception regardless of the identity of the perceived speaker. Therefore, while it remains possible that listeners may be more accurate at recognizing self-produced speech when they are aware that stimuli are self-produced, the current results challenge the common-coding prediction that self-produced speech will necessarily be recognized more accurately than the speech of others.

It is also important to consider why we do find advantages for the speech produced by the model speaker. For example, it may be the case that listeners assumed that the noise-vocoded speech was produced by an average talker, and this framing lead to facilitation in perception for the stimuli produced by the model speaker. This interpretation would be consistent with "talker normalization" (Nearey, 1989; Nusbaum and Magnuson, 1997), a theory which argues that listeners utilize information about the talker, and frame incoming speech with regard to the perceived identity of the talker, in order to overcome the considerable amount of uncertainty in speech and variability between speakers (Heald and Nusbaum, 2014). With regards to our experiment, assigning the identity of “average speaker" to the NVS stimuli would facilitate perception when the durational and amplitude properties of the stimuli fell closer to the statistical average of the community. However, given how abnormal NVS sounds (the stimuli were often described by our participants as sounding as if they were produced by a “demon" rather than by a normal person), it seems implausible that listeners identified the producer of this speech as an average speaker of their community.

We suggest instead that listeners may have either assumed that the NVS was not self-produced given how strange it sounded compared to the listeners’ own speech (or any other normal speaker) and thus must have been produced by some “other”, or failed to make any explicit assumptions (positive or negative) about the source of the NVS. Importantly, no matter what assumptions the listeners did or did not make, they appear to have adopted a “default” perceptual strategy, leading to facilitation in the identification of speech produced by the average speaker. This strategy may depend on the

lack of indexical information in NVS. For example, listeners have been found to generate more fine-grained expectations about the phonological qualities of upcoming words when these words are embedded in carrier sentences pronounced with the listener's own regional accent (Brunelliere and Soto-Faraco, 2013). However, when presented with a carrier sentence in a non-native regional accent, listeners do not make such fine-grained predictions. This demonstrates that listeners are able to draw upon more fine-grained linguistic representations to facilitate speech perception, but may only do so in the presence of indexical information cueing them to access such representations. In the context of visual action perception, the "default strategy" may involve accessing representations that more closely align with the participant's own production experience (Knoblich et al., 2002), leading to advantages for self-produced stimuli. Yet our results suggest that during speech perception, in the absence of vocal cues to identity, a video showing the participant's own face (Tye-Murray et al., 2013), or other indexical cues, the listener adopts a decoding strategy that relies on more general representations, and specifically not representations aligning to the listener's own productions.

This view complements talker normalization theory by suggesting that unless given evidence otherwise, the most efficient strategy during speech perception is to access representations that are more likely to match the input, that is, statistically average representations. However, given that linguistic and extra-linguistic cues to identity are almost always present to some extent during normal speech perception, it is important to consider how this may affect the interpretation of our findings. With regard to the implications for clear speech, it may be the case that when faced with unfamiliar speech, e.g. speech in an non-native language produced by an unfamiliar speaker, listeners access statistically average representations from their own linguistic community. Yet when additional information is available, indexical cues may "fine-tune" underlying representations (Brunelliere and Soto-Faraco, 2013), or cue the listener to utilize alternative processing strategies.

We hypothesized that if listeners utilize a "common-code" for both production and perception (Prinz, 1990), they should be more accurate at recognizing words based on their own recordings rather than those based on a speaker whose productions are more representative of the statistical average of the participant population, and that such effects should be found even in the absence of cues to the identity of the talker. While some participants were more accurate for self-produced stimuli than stimuli produced by the average speaker, the results of this study show an overall advantage for the stimuli produced by the average speaker, suggesting that spoken word recognition does not necessarily rely on common representations for production and perception. One of the advantages of a common-coding approach is that it eliminates the need to posit separate representations for speech sounds or word forms for both production and perception.

Yet in so doing, it complicates the question of how people perceive actions that differ from their own. As Hickok remarks, "...we are fully capable of understanding actions we have never produced...it would be surprising, maladaptive even, if all observed actions resulted in the activation of the same motor program in the observer" (2009). If comprehension critically relied on accessing the very same representations utilized in production, then we would expect that people who deviate greatly from the norm in their production would also suffer perceptually, yet this does not appear to be true (Scott et al., 2013). While self-advantages may be found for certain actions during visual perception, when faced with the task of recognizing words that vary across a number of speakers, the listener develops representations that extracts relevant information from different speakers and thus facilitates perception when faced with a novel speaker (Lively et al., 1993).

4.5 Conclusion

This study examined one facet of the debate on the nature of representations of speech, namely whether incoming speech is decoded with reference to representations that are more reflective of the overall speech community or more reflective of the listener's own productions. Our results suggest that, in the absence of indexical cues leading the listener to frame incoming noise-vocoded speech with respect to the identity of the speaker, the representations accessed during speech perception more closely align to the statistical average of the linguistic community rather than the speaker's own productions. This may suggest that indexical cues lead to fine-tuning of these underlying representations, but future research is needed to determine how and to what extent these results generalize to speech perception when such cues are present. The current results nevertheless already indicate that, contrary to a strict common-coding account, speech perception is not necessarily based on representations of one's own speech.

In interpreting our results, we emphasize the demands placed upon a listener to decode incoming speech; given a population in which different talkers may vary greatly with respect to the pronunciation of the same words or speech sounds, it is likely to be more efficient to attempt to decode incoming speech utilizing representations that closely approximate the statistical distribution of the community rather than one's own productions, which may deviate substantially from the average. While it is crucial to integrate research on speech production and perception, such research should take into account the differing demands on a speaker compared to a listener and how this may be reflected in the nature of the representations utilized for both types of language processing.

4.6 Acknowledgments

We thank two anonymous reviewers for constructive feedback.

4.7 Appendix: Word Lists

S1 Table: Word lists organized by Word Difficulty. Word is given in Dutch orthography. The second column contains a modified CPSAMPA transcription, which was utilized to calculate Levenshtein Distances. In the original CPSAMPA transcriptions, certain consonant sounds were designated by combinations of characters. In the modified versions, each consonant is designated by a single character. The third column refers to the lexical frequency of the word (as measured by Subtlex per-million). PTAN refers to the Number of phonological neighbors. PTAF refers to the average frequency of a word's phonological neighborhood. All information obtained from the Northwestern University DutchPOND database (http://clearpond.northwestern.edu/dutchpond.html).

Easy Word List

Word	Transcription	Translation	Freq_per_million	PTAN	PTAF
hotel	h.oU.t.E.l	hotel	88.7274	1	6.6546
onzin	O.n.z.I.n	nonsense	111.2295	1	0.8918
tafel	t.2.f.5.l	table	83.3992	3	2.4545
mens	m.E.n.s	man	144.6623	7	26.9384
schat	s.x.a.t	treasure	264.0328	17	39.9018
zelf	z.E.l.f	itself	437.4629	3	138.892
eind	Ei.n.t	end	83.1705	7	2.95
thuis	t.9y.s	home	399.4793	11	86.9666
agent	2.G.E.n.t	agent	186.5792	1	4.1162
gezin	G.5.z.I.n	family	79.8547	5	139.389
wapen	0.2.p.5.n	weapon	140.5918	11	78.8069
aarde	2.r.d.5	earth	100.0699	6	45.8692
geld	G.E.l.t	money	793.6076	11	17.6021
maand	m.2.n.t	month	93.644	15	60.2554
mama	m.2.m.2	mama	206.4514	1	1.6236
plan	p.l.a.n	plan	143.336	6	8.6364
schip	s.x.I.p	ship	115.2771	10	3.8464
stem	s.t.E.m	voice	86.5321	9	43.139
papa	p.2.p.2	papa	223.2822	2	6.6317

spel	s.p.E.l	game	95.1076	10	88.3661
stop	s.t.O.p	stop	301.9706	16	24.0842
auto	Vu.t.oU	car	457.9983	2	91.8946
kamer	k.2.m.5.r	room	275.238	7	14.9654
broer	b.r.u.r	brother	245.6241	4	30.6373
kaart	k.2.r.t	map	79.6718	20	17.4368
mond	m.O.n.t	mouth	165.9295	20	50.5346
fout	f.Vu.t	error	165.2663	18	53.4728
idee	i.d.eI	idea	482.9929	0	
wagen	0.2.G.5.n	car	77.8195	16	76.739
dame	d.2.m.5	lady	82.7361	3	56.6667
foto	f.oU.t.oU	photo	119.1646	2	0.5374
plek	p.l.E.k	place	177.4092	7	4.4625
soort	s.O3.r.t	kind	222.0244	14	43.2252
begin	b.5.G.I.n	beginning	193.1194	6	41.2994
geval	G.5.v.a.l	case	137.2074	5	28.562
geest	G.eI.s.t	spirit	94.9704	9	69.3914
paard	p.2.r.t	horse	83.6279	18	42.2472
prijs	p.r.Ei.s	price	86.6007	4	25.0403
vraag	v.r.2.x	question	436.6854	9	122.2365
twee	t.0.eI	two	1007.5139	5	13.0393
grond	G.r.O.n.t	ground	110.2461	5	40.3527
hulp	h.Y.l.p	help	239.7699	8	41.4652
spijt	s.p.Ei.t	regret	665.776	17	3.6642
stap	s.t.a.p	step	126.5281	17	53.3777
trek	t.r.E.k	pull	125.7277	10	13.2062
rest	r.E.s.t	rest	175.0309	18	43.2178
rust	r.Y.s.t	rest	75.3955	14	18.6341
gang	G.a.N	corridor	110.795	13	144.9526
derde	d.E.r.d.5	third	76.4931	1	24.6287
leger	l.eI.G.5.r	army	107.9822	8	9.9561
reden	r.eI.d.5.n	reason	163.6656	16	34.2018
kerk	k.E.r.k	church	79.4888	7	102.1117
stel	s.t.E.l	set	214.9354	17	64.7202
zorg	z.O.r.x	care	218.8229	2	28.8935
zaak	z.2.k	case	239.3354	15	95.6107
regel	r.eI.G.5.l	rule	77.3164	6	21.3396
kost	k.O.s.t	costs	80.4035	17	98.6239
nacht	n.a.x.t	night	204.439	13	148.8489

werk	0.E.r.k	work	680.2285	8	123.5495
kijk	k.Ei.k	look	1049.7738	18	21.5365

Hard Word List

Word	PhoWord	Translation	Freq_per_million	PTAN	PTAF
vaas	v.2.s	vase	4.5736	26	242.7665
vaat	v.2.t	vascular	1.4178	23	215.003
been	b.eI.n	leg	53.6252	18	566.7174
dier	d.i.r	animal	28.1046	25	657.8655
haan	h.2.n	cock	4.2763	18	647.2289
koor	k.oU.r	choir	6.8832	18	490.7083
thee	t.eI	tea	58.7934	19	742.9118
peer	p.I.r	pear	1.9895	21	216.9064
poen	p.u.n	moolah	11.7084	15	273.5062
haai	h.2.j	shark	9.4444	15	234.9067
teen	t.eI.n	toe	7.3863	20	305.4511
nier	n.i.r	kidney	3.9104	19	1461.1456
maan	m.2.n	moon	42.0998	21	1048.6892
hoed	h.u.t	hat	35.9483	31	1278.9684
hiel	h.i.l	heel	0.8461	17	373.9251
boor	b.oU.r	drill	3.4531	24	384.2521
tong	t.O.N	tongue	31.9007	16	188.3643
wijn	0.Ei.n	wine	60.4399	25	563.5474
touw	t.Vu	rope	26.2523	27	618.4513
cent	s.E.n.t	cent	36.9317	17	195.3484
halt	h.a.l.t	stop	17.4939	23	168.271
veer	v.I.r	spring	3.4302	19	239.2102
teer	t.I.r	tar	3.3387	20	227.802
zaal	z.2.l	room	15.413	19	153.8191
hout	h.Vu.t	wood	23.5768	22	1313.7525
maat	m.2.t	size	69.1754	38	666.6011
doel	d.u.l	goal	77.4536	18	333.4038
kier	k.i.r	crack	0.9833	20	245.8985
vouw	v.Vu	fold	1.7151	24	375.8025
maart	m.2.r.t	March	8.3925	22	437.2696
moer	m.u.r	nut	9.2158	17	742.6229
zaag	z.2.x	saw	3.5445	15	175.9578
goud	G.Vu.t	gold	61.9949	22	338.0655

mouw	m.Vu	sleeve	4.1848	29	678.966
tent	t.E.n.t	tent	40.9335	20	171.8877
wand	0.a.n.t	wall	3.293	28	463.0227
mand	m.a.n.t	basket	4.2992	25	173.9772
moes	m.u.s	puree	2.4011	17	284.0879
haas	h.2.s	hare	2.1725	33	317.3344
wang	0.a.N	cheek	7.8894	16	765.4
zool	z.oU.l	sole	0.9147	15	434.8941
bijl	b.Ei.l	ax	9.2615	17	175.955
gier	G.i.r	vulture	1.0062	16	305.911
meel	m.eI.l	flour	1.6922	17	233.6884
vacht	v.a.x.t	coat	3.3158	19	157.3781
moed	m.u.t	courage	41.0479	23	522.0363
noot	n.oU.t	note	3.8189	18	1224.6128
woud	0.Vu.t	forest	5.191	28	603.7991
lach	l.a.x	laugh	46.7191	20	173.158
zuil	z.9y.l	column	0.5488	15	158.6194
zeil	z.Ei.l	sail	6.9061	19	650.8264
boon	b.oU.n	bean	1.4178	19	341.0687
heil	h.Ei.l	salvation	5.717	20	520.0961
pauw	p.Vu	peacock	0.9605	25	334.1814
dauw	d.Vu	dew	0.9833	28	1482.717
graad	G.r.2.t	degree	4.0248	15	234.649
laan	l.2.n	avenue	1.2349	17	528.5336
maag	m.2.x	stomach	23.5539	17	610.9388
doos	d.oU.s	box	38.2809	23	188.8352
mijl	m.Ei.l	mile	15.3215	18	397.2535

Chapter 5

Contrasting ideomotor and general auditory accounts of speech intelligibility

Despite exhibiting wide variation in their vocal speech patterns, healthy listeners are often assumed to differ little with regard to their perceptual abilities (e.g., they will tend to agree on how intelligible they rate a given talker). This separation between production and perception is challenged by an ideomotor account that, in contrast to a general auditory account, argues for an influence of production experience on speech perception. We propose that these two standpoints can be operationalized with respect to two phonetic distance metrics: talker-listener similarity and talker prototypicality. In this study, we performed acoustic analyses on simple sentences recorded by native Dutch speakers. These same participants were then presented with degraded versions of the recorded sentences. Recognition accuracy for the degraded speech constituted a measure of intelligibility. We found that variation in talker intelligibility could be predicted by both talker prototypicality and talker-listener similarity. However, similarity effects were found to be much weaker than prototypicality effects and only held under certain conditions. These results may reflect the operation of both ideomotor and general auditory mechanisms that work in parallel to guide speech perception.

5.1 Introduction

Language exists in the unique usage of individual speakers. We refer to an individual's idiosyncratic style of producing speech as their "idiolect" (from *idio*, "personal", and *legein* "to speak"). Even within a small linguistic community, variation within and between speakers will endow speech with rich statistical variation. What may be true, yet is less readily apparent, is that in addition to their idiolect, each individual may also possess their own unique way of perceiving speech as well, their "idio-acusis" (from *akouein*, "to hear"). The need to coin such an unwieldy term reveals a tacit assumption that, while individuals within a linguistic community may differ in terms of their *produced* speech patterns, the representations utilized during speech *perception* are by and large the same. If a talker is considered to be very intelligible, we take that to mean that the talker is intelligible to most or all listeners, not only to specific individuals. Yet this assumption is untested.

If variation in experience affects the way we listen, both talker and listener variability will likely affect talker-listener intelligibility. This study utilizes acoustic analysis in addition to perceptual assessments in order to assess how the relationship between the production and the perception of speech might determine speech intelligibility. Specifically, we test two possible metrics that have been proposed to modulate talker-listener intelligibility: Similarity and prototypicality.

These two metrics stem from competing theories about the mechanisms and representations involved in perception; ideomotor accounts suggest that perceived events are mapped onto our own action experience, while general perceptual theories emphasize the statistical variation in the sensory input. We argue that these distinct standpoints can be operationalized, respectively, as the similarity of the talker's speech patterns to the listener's own speech patterns and as the prototypicality of the talker's speech patterns with respect to the speech community as a whole. These two theories can thus be translated into functional predictions about what properties will cause a given talker to be more or less intelligible to a particular listener.

Ideomotor theories of perception propose that an action, such as a hand movement, is coded in terms of its perceptual effects (James, 1890; Lotze, 1852; Shin et al., 2010). This refers not only to an action's immediate, "resident" effects coded in terms of visual and kinesthetic feedback but also to remote, action-contingent sensory events such as a change in ambient light caused by flicking a light switch (Hommel et al., 2001), a musical note produced by tapping a piano key. If an action is coded by reference to its sensory effects, it follows that at certain cognitive level there is no distinction between representations for production and representations for perception, that "action

and perception-related processes draw on identical cognitive codes..." (Schütz-Bosbach and Prinz, 2007, p. 351). Through repeated experience, an action and its sensory consequences become bound together such that the motoric program may be activated via perception or ideation of the sensory goal (Greenwald, 1970).

Action perception has been argued to involve direct matching (Iacoboni et al., 1999) or simulation (Jeannerod, 2001; Knoblich and Flach, 2001) of observed actions to corresponding motor programs within the perceiver. Similarly, there is evidence that the perception of speech may involve automatic simulation or mapping of the incoming speech signal onto one's own motor repertoire. Because these processes draw on the observer's own motor repertoire, perception is argued to be facilitated when the observed event resembles the observer's own sensorimotor experience, such as when one is viewing recordings of self-produced actions (Knoblich et al., 2002; Wilson and Knoblich, 2005). With regard to speech, participants have been found to be more accurate at lip-reading silent videos when watching videos of themselves as opposed to other participants (Tye-Murray et al., 2013). Recognition advantages for self-produced stimuli has also been found in audiovisual speech recognition under difficult listening conditions (Tye-Murray et al., 2015). These findings make a plausible case for a perceiver's sensorimotor experience guiding or strongly modulation perception, which in turn suggests that talker-listener intelligibility ought to be modulated by similarity of a given talker's speech patterns to those of the listener.

While ideomotor theory suggests a special role of production experience in shaping perception of conspecific actions, a general auditory theory posits no such differences. Rather, a listener maps the continuous acoustic speech signal onto abstract cognitive categories (Diehl et al., 2004; Holt et al., 2010), such as phonemes, syllables or lexical items, using the same mechanisms that categorize other environmental sounds, such as slamming doors (Fowler and Rosenblum, 1990). Furthermore, an ideomotor account of speech perception emphasizes the listener's experience as a speech *producer*, while a general auditory account emphasizes the listener's experience as *perceiver* of a range of idiosyncratic talkers.

Although speech categories have often been described phonologically as discrete categories composed of binary features (Chomsky and Halle, 1968; Jakobson et al., 1963), the acoustic input varies along continuous dimensions: "One can conceptualize a segment of the speech signal as a point in this space representing values across multiple acoustic dimensions" (Holt et al., 2010, p. 1218). Listeners have been found to be sensitive to variation along these continuous dimensions, which may possibly suggest that speech categories represent probability distributions (Clayards et al., 2008). For

example, when asked to judge the 'goodness' of phonetic exemplars (Massaro and Cohen, 1983) or when a phonetic categorization task is not a simple two-alternative forced choice, listeners demonstrate graded sensitivity to within-category variation (Gerrits and Schouten, 2004; Schouten et al., 2003). This suggests that representations for speech may bear greater resemblance to probability distributions (Clayards et al., 2008) than binary configurations.

Exposure to statistical variation in acoustic patterns has a profound effect on how auditory categories are shaped and how future speech sounds are categorized. For example, the well-documented ability of infants to discriminate non-native contrasts (Werker and Tees, 1984) is slowly weakened as infants are exposed more and more to the statistical regularities of their native language (Kuhl, 2004). If exposure leads to speech categories being mentally represented as a statistical distribution in a multidimensional space, the centroid of this space may act as a prototype of the category. Such prototypes may warp the perception of the speech signal (Feldman et al., 2009; Kuhl, 1991; Samuel, 1982), while individual exemplars can also be ranked according to their distance from this prototype (Grieser and Kuhl, 1989; Kuhl, 1989; Rosch, 1973; Rosch and Lloyd, 1978).

If speech prototypes are created by averaging input across a range of talkers, and speech perception makes reference to these prototypes, a talker who consistently produces acoustic exemplars that fall closer to the statistical average of the linguistic community (i.e., a prototypical talker) may be more intelligible than a less prototypical talker. In contrast with the self-advantages found for visual and audiovisual speech (Tye-Murray et al., 2013, 2015), single, noise-vocoded words produced by a statistically average speaker have been found to be more intelligible than words that were self-produced (Ch. 4). While it is unclear to what extent the perception of noise-vocoded speech is generalizable to non-degraded speech, this experiment suggests that talker prototypicality may be a more important component of intelligibility than talker-listener similarity. It is not the similarity of the talker to the listener that determines intelligibility, but rather the prototypicality of the talker with respect to the greater linguistic community.

Thus, ideomotor theory suggest that intelligibility is enhanced when the speech patterns of the talker and listener are similar to each other, while general auditory theories suggest that a more prototypical talker will be more intelligible regardless of their similarity to the listener. These predictions can be operationalized in terms of distance metrics. A talker can be represented as a set of values across a number of phonetic dimensions, and similarity between two talkers calculated as the distance from one talker to the other in this multidimensional space. Prototypicality can be calculated as the average distance from one talker to all other talkers in a sample.

In this experiment, we utilize phonetic analysis to obtain prototypicality and similarity measurements and then investigate to what extent these two metrics predict speech intelligibility. We recorded participants producing simple sentences containing two semantically unrelated key words (e.g., "the crown is above the fish"). In three subsequent sessions, we presented the same participants with acoustically degraded versions of the recorded sentences to obtain intelligibility measurements. Thus, participants listened to their own speech as well as the speech of other participants. We then utilized the distance metrics as independent variables to predict participant accuracy for different talkers and thus evaluate similarity and prototypicality as predictors of talker intelligibility.

In order to accurately measure the influence of these two distance metrics, it is important to control for additional factors known to influence speech perception. Stimuli were manipulated to create three listening conditions: Noise-vocoded speech (NVS; Shannon et al., 1995), speech in noise (SPIN), and speech in noise which had been filtered to approximate how people perceive themselves during self-monitoring (FiltSPIN; Vurma, 2014). These manipulations modulate the spectral properties of the signal to enable control over fine-grained spectral cues to talker identity, as well as control for spectral differences between recorded speech and actively produced speech.

With regard to bottom-up cues to speaker identity, previous experience listening to a specific talker is known to increase the intelligibility of novel speech produced by that talker (Dahan et al., 2008; Nygaard et al., 1994; Remez et al., 2011). Phonetic cues to identity may lead listeners to recognize a specific talker, leading to increased intelligibility. While listeners are sensitive to talker-specific variation in durational cues (Allen and Miller, 2004), conscious identification of a talker's identity on the basis of the auditory signal alone usually depends on fine-grained spectral cues (López et al., 2013). Utilizing NVS enables us to make talker identification from the acoustic signal extremely unlikely.

Top-down cues to talker identity, such as a label (Niedzielski, 1999) or a picture indicating the talker's race (Yi et al., 2013) may also influence how speech is perceived. To test whether a top-down cue to talker identity may modulate intelligibility, half of the participants were presented with a label identifying the talker prior to stimulus presentation. These manipulations and our acoustic analyses enabled us to test under various indexical conditions whether talker-listener intelligibility was more accurately predicted by similarity or by prototypicality. The experimental design also allows for comparison to previous experiments that found advantages for self-produced visual (Tye-Murray et al., 2013) and audiovisual (Tye-Murray et al., 2015) speech, due to the fact that participants listened to themselves and other talkers.

In addition to controlling for indexical cues that may modulate the perception of a talker's speech, it is also important to consider the difference between the sensory experience of producing speech and the sensory experience of listening to another talker. During speech production, our voices are filtered by both air as well as bone conduction (Békésy, 1949; Puria and Rosowski, 2012), modulating the acoustic signal. The cringe-inciting experience of listening to a recording of one's own voice stems in part from the discrepancy between the spectral properties of recorded speech compared to how we habitually hear our active productions. This mismatch may confound the testing of a similarity-based mechanism, for the common perception-action codes posited by ideomotor theory may correspond not to the spectral properties of a recording but to how speech sounds during self-monitoring. We take this into account by testing participants' recognition of unfiltered speech as well as speech that has been filtered in order to approximate the bone-and-air-conducted signal (Vurma, 2014).

We hypothesized that if speech recognition involves active comparison to one's own motor repertoire, or simulation of incoming speech using one's own production representations, then it is likely that similarity will be the best predictor of talker-listener intelligibility. However, if input from multiple talkers leads to the formation of abstract prototypes to which the incoming signal is compared, prototypicality ought to be a better predictor of intelligibility. It is important to note that, while contrastive, these predictors are not mutually exclusive; it is possible that intelligibility can be modulated by prototypicality, similarity, both, or neither. Indeed, any relevant knowledge or mechanisms that may enhance perception are likely to be deployed if possible (Poeppel et al., 2008).

5.2 Method

5.2.1 Participants

Forty-nine participants (21 men) were recruited from the Max Planck Institute for Psycholinguistics participant database for this experiment. All gave written consent before each session. Participants were paid for their time and informed that they were free to withdraw from the study at any time with no negative repercussions. The study was approved by the Ethics Committee of the Social Sciences Faculty of Radboud University.

Three participants (all female) dropped out of the study before moving on to the identification sessions, leaving 46 participants for the NVS session. A further three participants (two female, one male) failed to complete the SPIN and FiltSPIN session, leaving 43

TABLE 5.1: **Characteristics of lexical stimuli by Word Group.** Standard deviations are noted in parentheses. HFHP = High Frequency High Phonological Neighborhood Density, LFHP = Low Frequency Low Phonological Neighborhood Density, LFHP = Low Frequency High Phonological Neighborhood Density, LFLP = Low Frequency Low Phonological Neighborhood Density. Length refers to the number of phonemes in the spoken form of the word.

Group	Ave. Freq	max	min	Ave. PND	max	min
HFHP	345.24(888.05)	4025.84	12.65	22.71(7.37)	46	13
HFLP	38.21(46.58)	247.48	9.51	7.21(2.83)	12	2
LFHP	3.93(2.33)	8.48	0.55	21(5.12)	33	13
LFLP	3.63(2.24)	9.10	0.62	5.89(2.65)	11	1
Group	Ave. Length	max	min			
HFHP	3.21(0.41)	4	3			
HFLP	3.89(0.56)	5	2			
LFHP	3.11(0.41)	4	2			
LFLP	4.07(0.72)	5	2			

total participants for the FiltSPIN session. One male participant failed to complete the final SPIN session, leaving 42 total participants for the SPIN session.

5.2.2 Design

The experiment consisted of one recording session and three identification sessions (NVS, SPIN, and FiltSPIN). The study began with the recording session, followed by a rest interval of approximately two weeks, followed by the NVS session. The third and fourth sessions (either FiltSPIN or SPIN) began at least seven days after the NVS session. As the SPIN and FiltSPIN manipulations were very similar, these two sessions were separated by a rest period of two weeks in order to reduce learning effects on identification responses and the session order was counterbalanced across participants.

Participants were randomly assigned to small groups of seven same-gender talkers for the identification sessions; each participant was presented with an equal number of sentences, each containing two semantically unrelated target words, produced by each talker in their group (including the listener themself). This represented a compromise between having a single participant listen to a sufficient sample of stimuli from a single talker, while also being exposed to a range of different talkers. The entire sequence required approximately six weeks to complete.

5.2.3 Lexical Stimuli

Previous research has demonstrated that under adverse listening conditions, listeners correctly identify higher frequency words with few phonological neighbors more often than low frequency words with many phonological neighbors (Bradlow and Pisoni, 1999). This suggests that an adequate comparison across talkers would be biased by the recognition difficulty of the lexical stimuli. In order to ensure that words of varying difficulty were equally allocated across talkers, we collected 112 Dutch words and sorted them into four categories (Word Groups) based on lexical frequency and phonological neighborhood density. We define a phonological neighbor as any existing word which differs from a source word by the substitution of a single phoneme. Details of these words groups are given in Table 1. To elicit the utterances, line drawings of the referent objects were selected from a picture data base available at the Max Planck Institute for Psycholinguistics.

Lexical stimuli were embedded in sentences. Each sentence contained two target words, e.g., “kroon” and “vis”, and described the two objects as having a simple spatial configuration, as in “de kroon is boven/onder/naast de vis” (“the crown is above/under/next to the fish”). Each content word was preceded by the definite determiner “de” (“the”).

5.2.4 Recording Procedure

Participants were seated in a sound-attenuated booth approximately 3 to 5 centimeters away from a pop-filter shielded microphone (Sennheiser ME 64). They were first asked to say several words, such as counting to 10, in order to habituate to the acoustics of the recording booth. Following acclimation, the recording level was set individually according to each participant’s speaking amplitude. Due to the fact that read speech differs from spontaneous speech in terms of intonational, durational and spectral features, and that listeners are sensitive to these differences (Howell and Kadi-Hanifi, 1991; Laan, 1997), it seemed likely that read materials would be limited in their generalizability to natural speech. However, the demands of the identification task required strict control over the format of the sentences. Therefore sentences were elicited in a “semi-spontaneous” format; while the exact form of the sentence was pre-specified on each recording trial, at no time did participants simply read words off of a screen.

A trial began with a fixation cross. After 200ms, the fixation cross disappeared and the orthographic form of the first target noun appeared on the screen for 500ms. This was followed by a blank screen for 200ms, after which the second noun appeared for another 500ms. After presentation of another blank screen for 200ms, two line drawings

depicting the target nouns appeared on screen arranged either horizontally or vertically. The order of orthographic word presentation determined the spoken order of the sentence. For example, if the display consisted of line drawings of a witch in the bottom half of the screen and a ball on the top half of the screen, and the order of word presentation had been "witch – ball", the participant would then say "the witch is under the ball." All drawings were taken from the Max Planck Institute for Psycholinguistics picture database.

The recording session began with 10 practice trials using the same two content nouns, "zebra" and "penguin", with different spatial configurations ("above", "below", "next to") and different word orders (e.g., "the zebra is above the penguin", "the penguin is above the zebra") in order to acclimatize participants to the task. Participants were instructed to try to speak as naturally as possible. Each of the 112 target nouns appeared in both initial and final position, comprising 112 sentences. Three lists were generated and participants read all three lists, comprising 336 total sentences. The order of lists and sentences within lists was fully randomized for each participant. A researcher listened to each production to ensure that participants correctly produced the target sentence, that it was produced in a natural manner, and that each recording was free of coughs, stutters, or long gaps which could compromise the intelligibility of the target words or potentially be utilized by a listener as a cue to talker identity. Because the experiment design required that all participants produced all stimuli, trials with such mistakes were repeated until an acceptable version had been recorded. Recordings containing mistakes were discarded and not used for analysis. The entire recording procedure took approximately one hour.

5.2.5 Stimulus Preparation and Manipulation of Spectral Indexical Information

Within each group of seven same-gender talkers, all recordings were normalized to root-mean-squared amplitude. Leading and following silence in each recording was standardized to 500ms. For each identification session, the recordings were manipulated in order to modulate the fine-grained spectral information present in the signal. These three manipulations were noise-vocoding (NVS), speech-in-noise, and speech-in-noise that had been filtered to approximate how a person hears their own voice when they are speaking (FiltSPIN). While all manipulations increase the difficulty of word recognition, NVS eliminates almost all spectral cues to talker identity, while such cues are preserved in SPIN and FiltSPIN.

5.2.5.1 Noise-Vocoded Speech

Noise-vocoding is a method utilized to systematically degrade speech, often in order to simulate the type of signal perceived via cochlear implant (Davis et al., 2005; Shannon et al., 1995). First, the acoustic spectrogram (which specifies the amplitude of different frequencies in the signal as a function of time) is separated into a number of frequency bands. At each point in time, the average amplitude of all frequencies within a given band is extracted and these values are then utilized to modulate the energy of broad-spectrum noise in the corresponding frequency bands. Varying the number of bands can systematically modulate the intelligibility of the original signal (e.g., Sohoglu et al., 2014).

For this experiment, we separated the signal into six frequency bands corresponding to equally spaced distances along the basilar membrane (Greenwood, 1990). Due to the averaging process, fine-grained spectral cues to speaker identity are destroyed, leaving only amplitude and duration cues. Anecdotal evidence from the current and previous experiments (Ch. 4) suggests that most listeners are unable to distinguish between talkers in noise-vocoded speech, except in the case of salient and memorable cues to duration. In this condition, therefore, it is extremely unlikely that participants were reliably able to identify the identity of a talker from the properties of the acoustic signal (when no label was present).

5.2.5.2 Speech in Noise

Speech in noise (SPIN) is a standard technique in which a speech signal is embedded in speech shaped noise at a specified signal-to-noise ratio (SNR). A decrease in the amplitude of the signal relative to the embedded noise increases the difficulty of word recognition. For this study, we utilized an SNR of -7 decibels, which in pilot testing was found to be comparable in difficulty to 6-band NVS. However, unlike in the NVS session, all fine-grained spectral cues to talker identity remain intact with this manipulation.

5.2.5.3 Filtered Speech in Noise

In addition to the SPIN session, we also presented participants with speech in noise which had been filtered in order to simulate what is heard during self-monitoring (Filt-SPIN). Under normal speaking conditions, the speech signal is altered by the effects of bone-conduction (Békésy, 1949; Puria and Rosowski, 2012) as well as the stapedius reflex (Sesterhenn and Breuninger, 1978). In order to best approximate the speech signal

heard during self-monitoring, we utilized a filter that takes both of these processes into account (Vurma, 2014). This manipulation maintained the fine-grained spectral cues important for talker identification, yet transformed the recordings into a format that more closely approximates the experience one has when producing speech.

5.2.6 Identification Task

In each of the three identification sessions, participants listened to one type of the prepared materials and performed a two-word identification task. Given that the structure of the stimulus sentences was highly predictable, we asked participants to attempt to recognize only the two target nouns.

Participants were seated in a sound-attenuated booth in front of a keyboard and a computer screen. During each identification session, participants were presented with manipulated versions of the recordings and asked to identify the two target nouns in each sentence by typing in their responses via keyboard. Each trial began with the appearance of a fixation cross on the screen. Stimulus presentation began after 500 milliseconds. Following stimulus presentation, the fixation cross disappeared and the words "Eerste woord intypen. Druk op [ENTER] om te bevestigen" ("Type in the first word. Press [Enter] to confirm") appeared at the bottom of the screen. After responses for both words had been registered, the typing screen was replaced by a fixation cross and the following trial began.

Every session began with 10 practice trials, consisting of manipulated versions of the same sentences used during the practice phase of the recording session (recorded by a female native speaker of Dutch). The predictability of the practice stimuli gave participants a short amount of time to habituate to the properties of the stimulus manipulation. Following the practice phase, the test phase began. The stimuli were presented within the same lists that were used during recording in order to separate repetitions of the same word in the same position. The order of lists and order of stimuli within lists was fully randomized for each participant.

As all 336 sentences had been recorded by all talkers, it was possible to randomly select a talker for a given sentence on each trial. At least one sentence from all possible talkers (seven per group) was presented before repetition, thus the same talker could only appear a maximum of twice in a row. In order to balance the experimental conditions across talkers, assignment of sentences to talkers was pseudo-randomized such that an equal number of lexical items, both in initial and final position, were drawn from each of the four groups. Thus, for a given session, one talker provided 48 sentences (96 target

words), with 12 words from each word group in initial position and 12 words from each word group in final position.

5.2.7 Manipulation of Top-Down Cues to Talker Identity

For twenty-four participants (10 men), a top-down cue to talker identity was provided on each trial in place of the fixation cross. This cue consisted of a simple orthographic label written in lowercase letters denoting the "name" of the talker, which stayed on screen for 500ms. The label was removed prior to onset of the auditory stimulus. When the sentence had been produced by the same participant performing the identification task, the label was set to the second person singular pronoun "jij" ("you"). For the other six talkers, each talker was randomly assigned a name at the beginning of each session. Names were drawn from the six most popular male and female Dutch names given to persons born between 1990 and 1994, as recorded by the Netwerk Naamkunde project hosted by the Meertens Instituut (Netwerk Naamkunde, 2015). Specifically, female names were Laura, Maria, Anne, Lisa, Michelle, Iris, and Sanne; male names were Kevin, Thomas, Jeroen, Johannes, Tim, and Dennis. One female participant had the same name as one of the labels, though this was not thought to constitute a problem as in everyday life many people share the same name and the use of the second person pronoun adequately distinguished between talkers.

5.2.8 Phonetic Talker Comparison

For the purposes of quantifying each talker's speech in terms of prototypicality and similarity, we conducted an acoustic analysis of the recordings. Each participant produced 336 sentences, comprising 16464 total recordings. Each target noun was produced three times in both sentence initial and sentence final position. Initially, we attempted to extract as many potentially relevant acoustic variables from the data as possible. With regard to variation at the level of the entire utterance, for each recording we calculated sentence duration, average pitch/fundamental frequency (F0), range F0, average amplitude, and range amplitude.

Acoustic information at the level of individual phonemes within words was also collected. Measurements for all segments (vowels and consonants) included average duration, average F0, as well as amplitude measures in the form of mel-frequency cepstral coefficients (MFCCs; Davis and Mermelstein, 1980). MFCCs use filters to quantify amplitude within specific frequency bands in addition to the overall amplitude of the signal,

and are widely employed in computer-based speech recognition. Additionally, for vocalic segments, we measured averages of the resonant frequencies of the vocal tract (formants one to three). For any given segment produced by a specific talker, e.g., the vowel /e/, this segment can be quantified as a point along multiple acoustic dimensions (duration, amplitude, first formant frequency, etc...).

Due to the large number of sound files, automated forced alignment was utilized to identify segment boundaries. This introduced an additional level of complexity, given the fact that automatic alignment is not 100% accurate (though, of course, human aligners may also differ in their alignment judgments; see Goldman, 2011). This suggests that for any single token of a target segment, automatic alignment may fail to mark the relevant boundaries. However, the presence of such occasional errors does not preclude the possibility that reasonably accurate estimates of overall averages can be estimated based on automatically obtained alignments. Given a sufficiently large number of instances and adequate outlier detection, it is indeed possible to arrive at reasonable estimates for different variables, e.g., the average duration of the vowel [i]. Furthermore, using forced alignment does not rule out the possibility of reasonably estimating parameters typically derived from segment centers, such as spectral measurements, even when the actual segmental alignment is not 100% accurate. With adequate data cleaning and a sufficiently large sample size, automatic alignment can be used to quantify average values for different talker-specific acoustic variables. Details of this procedure are given in Appendix 1.

5.2.8.1 Phonetic Variable Selection

Having extracted reliable phonetic measurements from the recordings, the next step involved calculating measurements of similarity and prototypicality. We take the variable "F1" as an example. Each talker has an average value for F1 for a given segment, such as [ɛ]. If we center these scores by subtracting the mean value across all talkers (and standardize by dividing by the standard deviation), the resulting value quantifies for each talker how different their individual value of F1 is for that segment from the average of the community. A talker with a value closer to zero therefore has a more prototypical F1 for that segment than a talker with a larger value. If we repeat this standardization across all segments, we create an n-dimensional space. We can then calculate the Manhattan distance, a distance metric similar to Euclidean distance but more applicable to high-dimensional data, between all talkers in this space to create a measure of how much each talker differs from each other talker over all segments.

Talker-listener distance scores can be seen as a quantified measure of talker similarity (smaller distances correspond to phonetically similar talkers). For each individual talker, we can then take the average of these distances to other talkers. If a talker is very similar to many other talkers, this average will be low, and the talker can be said to produce more prototypical values for F1 than another talker whose average distance score is very high. This procedure was carried out for all variables. Given that participants only listened to same-gender talkers, distances were calculated for the male and female participants separately.

This procedure provides us with a number of candidate predictor variables consisting of average distance measurements between talkers for a given acoustic variable. Due to the wide range of acoustic factors that speakers may differ on, many studies that measure talker similarity (e.g., Bartoli et al., 2015) or assess talker convergence (Babel and Bulatov, 2011; Pardo, 2006) utilize listener similarity judgments rather than acoustic measures or a combination of acoustic and perceptual judgments (Pardo, 2013; Pardo et al., 2013). However, these methods are specifically designed to measure the overall similarity of two talkers, rather than similarity with regard to acoustic variables that are important for determining the intelligibility of a talker's speech. Therefore, if we utilize all variables in calculating prototypicality, we assume that all variables are equally important in determining the prototypicality of the talker, e.g., that having an atypical duration for the vowel [ɛ] affects intelligibility to the same extent as having an atypical pitch value for that same vowel. We were specifically interested in selecting acoustic variables that modulate the intelligibility of speech in the three identification sessions. We therefore utilized random forest analysis (Breiman, 2001) to determine the relative importance of the candidate variables for explaining variance in the identification results and then excluded non-important variables from analysis. That is, we attempt to predict variation in the identification task utilizing prototypicality and similarity measures derived from specific acoustic variables in order to determine which acoustic variables may be important for accounting for changes in intelligibility.

Random forests are useful for examining datasets in which there are a large number of (correlated) predictors and a small number of observations, such as in the case of selecting relevant genes for the classification of microarray data (Cutler and Stevens, 2006; Díaz-Uriarte and Alvarez de Andrés, 2006). The success of the predictors in accounting for variance in the outcome variable, based on a "forest" of classification and regression trees (CARTs), quantifies for each independent variable its 'importance' as a predictor. The use of random forest models is becoming more prevalent in the analysis of linguistic data (for an example of such an analysis and a clear, in-depth explanation of this method, see Tagliamonte and Baayen, 2012). For example, random forests have been recently used to examine the relative importance of psycholinguistic and turn-taking

variables for explaining the duration of floor transfer offsets in telephone conversations (Roberts et al., 2015).

All analyses were implemented using the open-source statistical software R (R Development Core Team, 2013). Random forest models were fit separately for the data from each identification session, using the *cforest* function from the package *party* (Strobl et al., 2009). For each talker's recordings, we calculated the average number (across all experimental conditions) of accurate responses for the six other participants in the talker's group. Crucially, we excluded data in which the participant and the talker were identical, i.e., when participants were listening to themselves. This comprised an intelligibility measure for each talker that served as the outcome variable for the random forest model.

A random forest model was run containing 1000 trees and 5 variables in each tree. Due to the fact that the data contained a high number of correlated predictors (especially for the computed predictors), we computed conditional variable importance and set the threshold for retention as the absolute smallest value of variable importance (Strobl et al., 2007). While multiple runs of the model with different starting seeds produced highly correlated importance scores (all $r > 0.95$), there were less important variables that only crossed this threshold on certain runs. To be conservative, we therefore ran each random forest model three times with different starting seeds and included only variables that passed the inclusion threshold on all iterations. For the NVS session, the phonetic predictors that mattered most for intelligibility were average sentence duration, average segment amplitude in MFCCs 0, 2, and 8, as well as average segment F1 frequency and average segment F1-F0 dispersion. For the SPIN session variables included average sentence duration, MFCC 10, F2 and F2-F0 dispersion. Similarly, variables selected for FiltSPIN session were sentence duration, MFCC 3, MFCC 10, F1, F2, and F2-F0 dispersion.

A similar method was utilized to obtain a measure of similarity between talkers. Instead of measuring intelligibility as average accuracy across participants when listening to a given talker, we took each participant's average accuracy for each specific talker as our metric of talker-listener intelligibility. However, instead of averaging over all distance scores, we measured distance between individual talkers. In contrast to the random forest analysis for the prototypicality scores, cases in which participants were listening to their own recordings (i.e., where the distance is equal to zero) were included during the variable importance calculations. As with the variable selection for prototypicality, some variables failed to reach the importance threshold over multiple runs with different starting seeds. Again, we ran each model three times and excluded all variables that did not pass the inclusion criterion in every run. For NVS, variables exceeding the threshold for

inclusion were average sentence duration, average sentence F0, MFCCs 0,1,2,3,4,9,11, F0, F1, F2, and F1-F0 dispersion. For the SPIN session, variables included average sentence duration, average sentence amplitude, average segment duration, MFCCs 1, 5, 8, and 9, as well as F2, F3, and F2-F0 dispersion. For the FiltSPIN session, variables included average sentence duration, average segment duration, MFCCs 2 and 3, as well as average F2, F3, and F2-F0.

Having quantified talker-listener similarity and talker prototypicality for each of the candidate phonetic variables, the random forest models enabled us to determine which specific variables appear to be important for explaining variance in talker intelligibility. While it is interesting to consider what specific variables are deemed important by a model, determining which parameters are important for intelligibility under these different degradation manipulations is not the focus of this experiment. Rather, we utilized this method in order to determine whether participants whose speech is more prototypical (with regard to phonetic variables important for intelligibility) are more intelligible than other participants when producing the same lexical items. Therefore, following variable selection, we recomputed distance scores, beginning with all individual measurements for the variables deemed important in the random forest analysis. This procedure generated session-specific measures of talker prototypicality and talker listener similarity based only on acoustic variables likely to be important for determining talker intelligibility in the identification task.

5.3 Results

In order to compare the typed-in responses to the auditorily-presented target words, all target words and participant responses were broadly transcribed into DISC, a computer readable phonetic orthography. This eliminated the influence of orthographic variation with no phonetic realization (e.g., in Dutch, word final “t” and “d” are both realized as [t]). An identification response was marked as correct (1) if the response transcription and the target transcription were identical, and incorrect (0)otherwise.

5.3.1 Experimental Factors

Accuracy was fairly consistent across sessions, averaging 54.5% in the NVS session and 55.3% in the SPIN session, with slightly lower accuracy in the FiltSPIN session (49.9%). When the position of the word (First or Second) was taken into account, larger differences emerge between sessions. In the NVS session, accuracy was greater for the second

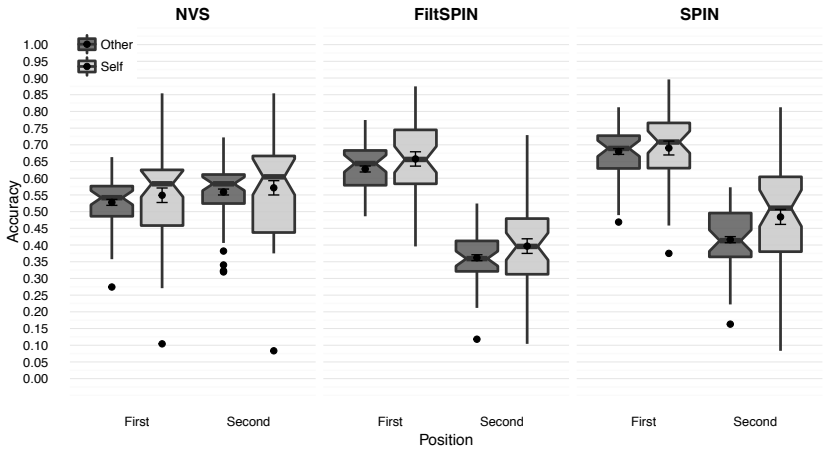


FIGURE 5.1: **Average and distribution of accuracy in NVS, FiltSPIN, and SPIN sessions by Position and Talker Condition.** Group averages indicated by the black dots (with error bars) within the boxplots, while by-participant average accuracy in each combination of conditions is depicted by the boxplots themselves. Dark gray fill refers to stimuli averaged across the six other talkers presented to each participant, while light gray fill refers to self-produced stimuli.

word (56%) than the first (53.1%). However, in both the SPIN and the FiltSPIN session, accuracy was much greater for the first word (68.1% and 63.2%, respectively) than for the second word (42.6% and 36.7%). Many participants reported that the word seemed to "fall off" towards the end of the sentence, though the phonetic data indicates that the amplitude of the second half of the sentence did not differ from the first (first half = 56.33 dB, second half = 56.62 dB). Given that the speech-shaped noise was calculated based on the average amplitude of the file, it may be that slight variations in speaking amplitude lead to slight deviations from the target signal-to-noise over the course of the sentence. Accuracy appeared to be slightly greater when a label was present (NVS: Label = 56.7%, No Label = 52.7%; SPIN: Label = 56.8%, No Label = 53.9%; FiltSPIN: Label = 50.9%, No Label = 52.7%). Similarly, accuracy appeared to be slightly greater when stimuli had been self-produced rather than produced by other participants (NVS: Self = 56%, Other = 54.3%; SPIN: Self = 58.7%, Other = 54.8%; FiltSPIN: Self = 52.7%, Other = 49.5%). Accuracy for the different combinations of Talker Condition and Position, for each session, are given in Figure 5.1.

Identification responses were analyzed using logistic mixed effect regression in R with the lme4 package (Bates et al., 2015). In order to reduce model complexity, separate models were fit for each session (NVS, SPIN, FiltSPIN). Fixed effects for all models included three within-participant variables (Word Group, Position, Talker Condition), as well as one between-participant variable (Label). Word Group consisted of four levels (see Table 1). Label consisted of two levels (Label/NoLabel), as did Position (First/Second). Talker Condition also consisted of two conditions: Self, when participants were

listening to self-produced stimuli, and Other, collapsed over all cases in which participants were not listening to self-produced stimuli. Significance of a fixed effect term was determined by maximum likelihood comparison of nested models with and without the target term.

Simulations by Barr et al. (2013) indicate that models containing only random intercepts suffer from inflated Type 1 error and the authors suggest utilizing a maximal random effects structure. Given the large number of categorical factors in this analysis, it would be unfeasible to utilize a fully specified random effect structure containing random slopes for all appropriate main effects and interactions. Therefore initial model fitting utilized only random slopes for main effects, excluding interactions. In cases where an interaction term was shown to be significant, we refit the model with an alternative random effect structure containing random slopes for the target interaction term, and utilized this model for assessing whether removing the interaction term indeed significantly affected model fit. All final models included crossed random intercepts for Participant and Word, as well as random slopes for Position, Talker Condition and Word Group over Participant, and random slopes for Position, Label and Talker Condition over Word.

For the NVS model, only removal of the fixed effect for Position was found to significantly decrease model fit ($\chi^2(1) = 7.241, p < 0.008$). Model estimates indicate a slightly higher probability of an accurate response for words in second position ($\beta = 0.14, SE = 0.053$) compared the first. This is reflected in the slightly greater proportion of accurate responses for words in the second position in the sentence ($mean = 0.561$) than for the first word in the sentence ($mean = 0.531$).

For the SPIN model, removing the fixed effect of position also significantly decreased model fit ($\chi^2(1) = 120.45, p < 0.001$). In contrast to the NVS session, accuracy was lower for words in second position ($mean = 0.425, \beta = -1.274, SE = 0.064$) compared to the first position ($mean = 0.681$). Removal of the fixed effect of Talker Condition also significantly decreased model fit in the SPIN session ($\chi^2(1) = 7.955, p < 0.005$). An accurate response was slightly more likely when a participant was listening to a self-produced word ($\beta = 0.22, SE = 0.076$) than words produced by other talkers, as reflected in the difference in the proportion of accurate responses for words in the Self condition ($mean = 0.587$) compared to the Other condition ($mean = 0.548$).

For the FiltSPIN model, again the main effects of Position ($\chi^2(1) = 118.14, p < 0.001$) and Talker Condition ($\chi^2(1) = 0.024$) were found to significantly influence model fit. Under the simplified random effect structure, an initial interaction was found between Position and Talker Condition, however this did not survive model comparison of nested models that included a random slope for this interaction term. Accurate responses were

less likely in second position ($mean = 0.367, \beta = -1.274, SE = 0.065$) than in the first position ($mean = 0.632$), and more likely when words were self-produced ($mean = 0.527, \beta = 0.174, SE = 0.075$) than when produced by other talkers ($mean = 0.494$).

The results from the experimental factors suggest that accuracy is indeed slightly greater for self-produced stimuli compared to stimuli produced by other talkers, though this effect is not found when fine-grained spectral cues are destroyed (NVS session). Contrary to Chapter 4 , none of the models revealed any significant effects for Word Group. This may be due to the fact that the current experiment involved listening to seven talkers with a maximum of two consecutive repetitions of the same talker, rather than a blocked structure with only two talkers. This design may have decreased the importance of lexical characteristics compared to phonetic variation in determining word recognition.

The lack of a significant effect of Label, or a significant interaction between Label and Talker Condition, suggests that the modest self-advantage does not appear to depend on conscious knowledge that one is listening to one's own speech (as argued by Knoblich et al., 2002). Having established the significant experimental variables in first stage of model fitting, we then examined whether including additional measures of phonetic prototypicality and phonetic similarity improved model fit.

5.3.1.1 Phonetic Prototypicality and Similarity

Similarity and prototypicality were operationalized as distance metrics, such that more positive values indicated greater distance between a talker and a listener (in the case of similarity) or greater average distance from one talker to all other same-gender talkers (and thus greater distance from prototype). Prior to standardization, similarity distance scores for cases in which the talker is listening to their own stimuli are equal to 0. Prototypicality is calculated excluding cases in which listeners were presented with self-produced stimuli. The distribution of similarity and prototypicality scores is given in Figure 5.2. Inspection of variance inflation factors indicated no problems with multicollinearity between Prototypicality and Similarity.

In order to facilitate model convergence, the raw prototypicality and similarity values were first centered and scaled. These numerical covariates, as well as the interaction term between them, were entered into the models containing the significant experimental variables. Backwards-fitting and maximum likelihood comparison was utilized to assess the significance of the similarity scores. Random effect structures were specified for each model individually.

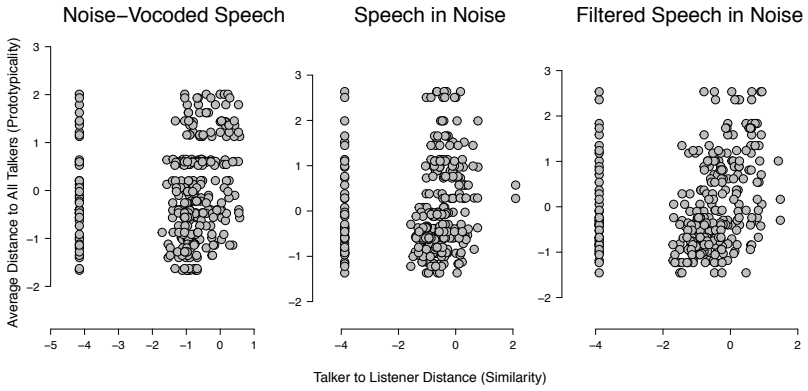


FIGURE 5.2: **Phonetic Prototypicality and Phonetic Similarity.** Similarity (x-axis) represents the distance between a talker and a listener based on the phonetic variables obtained from the random forest analysis. The vertical band of data points on the left-hand side of each graph represents cases in which the listener is identical to the talker (distance = 0 prior to standardization). Prototypicality (y-axis) represents the average distance from one talker to all other same-gender talkers. Greater values indicate that a talker is *less* similar to the listener or less prototypical, respectively. All scores have been standardized.

For the model for the NVS session, the random effect structure consisted of random intercepts for Participant and Word, random slopes for Position, Similarity and Prototypicality over Participant, and random slopes for Position and Similarity over Word (models with a random slope for Prototypicality failed to converge). Removing the interaction term between Similarity and Prototypicality as well as the main effect of Similarity failed to significantly affect model fit. Removal of Prototypicality did significantly decrease model fit ($\chi^2(1) = 41.942, p < 0.001$). Model estimates indicated a lower likelihood of a word being correctly recognized if the talker was less prototypical ($\beta = -0.17, SE = 0.02$).

The SPIN random effect structure also consisted of random intercepts for Participant and Word, random slopes for Position, Similarity and Prototypicality over Participant, and random slopes for Position and Similarity over Word. Model fit was not affected by removal of the three-way or two-way interaction terms. Removal of the main effects of Position ($\chi^2(1) = 118.09, p < 0.001$), Prototypicality ($\chi^2(1) = 56.79, p < 0.001$) and Similarity ($\chi^2(1) = 6.6203, p < 0.0101$) all significantly reduced model fit. Model estimates for position indicate that an accurate response was less likely when a word appeared in second position ($\beta = -1.27, SE = 0.06$). The probability of a word being accurately recognized was lower not only when the talker was less prototypical ($\beta = -0.21, SE = 0.18$), but also when the talker was phonetically less similar to the listener, though the magnitude of the effect is much smaller compared to Prototypicality ($\beta = -0.05, SE = 0.019$).

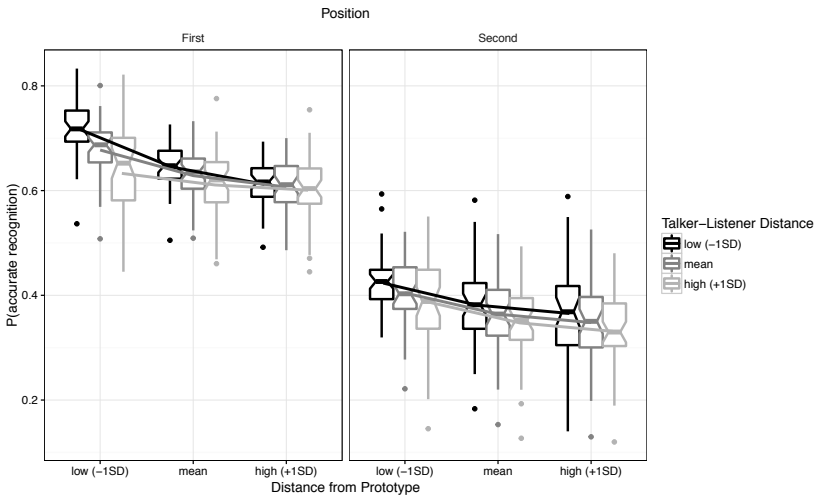


FIGURE 5.3: **Predicted probabilities of accurate responses in FiltSPIN model for different combinations of Prototypicality (average distance from all other talkers) and Similarity (talker-listener distance).** Lines indicate average predicted values, while boxplots indicate distribution of predicted values accounting for the by-participant random intercepts of the mixed-effects model. When a word is first position, a talker who is both highly prototypical (low average distance) and highly similar to the listener (low talker-listener distance) will be the most intelligible, but if the talker is aprototypical, then there is little to no benefit from low talker-listener distance. When a word is in second position, however, the effects of similarity are strongest when a talker is aprototypical and weakest when a talker is highly prototypical.

For the FiltSPIN session, the random effect structure initially contained random intercepts for Participant and Item, as well as random slopes for Position, Prototypicality and Similarity over both Participant and Item. However, model comparison suggested a significant three-way interaction between Position, Prototypicality and Similarity. In order to determine whether this significant interaction was not simply due to the lack of a random slope for this term (Barr et al., 2013), we refit the models including random slopes for main effects as well as two- and three-way interaction terms over Participant. Due to model fitting constraints, by-Word random slopes consisted of main effects only. Model comparison confirmed the significance of the three way interaction ($\chi^2(1) = 8.086, p < 0.005$). Inspection of the model parameters suggests that, as in the SPIN model, the probability of a word being recognized correctly decreased when talkers were less prototypical ($\beta = -0.17, SE = 0.045$) or less similar to the listener ($\beta = -0.06, SE = 0.032$). The estimates for the interaction between Prototypicality and Similarity ($\beta = 0.104, SE = 0.033$) suggest that the effects of talker-listener similarity are diminished when the talker is less prototypical. However, the estimates for the three-way interaction with Position ($\beta = -0.07, SE = 0.036$) suggest that this reduction of the effect of Similarity only holds when words are in first position, while in second position, stronger effects of talker-listener similarity are found when a talker is less prototypical (Fig. 5.3).

Data were also analyzed excluding trials in which participants listened to self-produced stimuli. Variable importance and distance measures were also recalculated excluding such trials. Using these alternative methods, the three-way interaction for the SPIN model approached significance ($\chi^2(1) = 3.576, p = 0.06$), and the three-way interaction for FiltSPIN was only significant when including a by-Participant random slope for this interaction. In contrast to the FiltSPIN interaction plotted in Figure 5.3, the predicted interaction between Similarity and Prototypicality is reversed for words in second position. That is, when excluding Self trials from all stages of analysis, the models predict that when a word is in second position and a talker is aprototypical, that talker will be more intelligible if they are *less* similar to the listener. Interestingly, model comparison for the NVS session without Self trials also suggested a three-way interaction between Position, Similarity, and Prototypicality. Estimates indicate suggest that the effects of similarity (that intelligibility is worse when the listener and talker are dissimilar) are *reversed* when a word is in first position and the talker is very prototypical (i.e., a talker who is very dissimilar will be more intelligible). However, all random effect structures including by-Participant random slopes for this three way interaction failed to converge, which calls into question the validity of this interaction. Furthermore, using an alternative model fit with a simplified random-effect structure, parameter estimates for main effects did not differ from those reported in the NVS model that included Self-trials. We restrict our discussion to results obtained from models including Self-trials.

5.3.2 Discussion

We investigated to what extent the intelligibility of a talker's speech would be better accounted for by the phonetic prototypicality of that speech or by the phonetic similarity of the talker's speech to that of the listener. Specific experimental factors were introduced in order to control for lexical variation in the stimuli as well as variation in indexical cues to talker-identity. The design of the study also enabled us to investigate whether participants exhibit an advantage for perceiving self-produced speech compared to the speech of other participants.

In accordance with previous findings (Tye-Murray et al., 2013, 2015), we did find a self-advantage for word recognition. However, this advantage was only found in the SPIN and FiltSPIN session, in which fine-grained spectral cues to talker identity were present. At first, this might lead one to conclude that such self-advantages depend on conscious recognition of one's own voice. Yet, if this were so, we would expect to have found an advantage for self-produced speech in the NVS session when each talker was accompanied by a label (thus ensuring that each participant knew who was speaking). In contrast, we found no self-advantage in the NVS condition, regardless

of whether or not self-produced speech was identified as such by the presence of a label. Top-down knowledge about who is speaking was not sufficient to obtain such self-advantages; increased intelligibility required the availability of fine-grained spectral cues in the signal.

While having received less attention, "other-advantages" have also been found in action perception. Zhu and Bingham (2014) contrasted common coding (Prinz, 1990) accounts of action perception with information-based theories of the perception of kinematics (Runeson and Frykholm, 1983). Participants with varying levels of motor expertise with regard to the perceived action watched point-light displays of actors throwing a ball towards one of nine spatial locations. In contrast to a similar study that found self-advantages for predicting where a thrown dart would land (Knoblich and Flach, 2001), Zhu and Bingham (2014) found that viewers were consistently *worse* at judging self-generated stimuli (Exp. 1), and that motor expertise in a gender-specific repertoire (i.e., underhand or overhand throw for softball and baseball respectively) did not modulate perception of same/other repertoire actions. In this regard, it may be that converting video into a point-light display and noise-vocoding audio recordings can be seen as congruent manipulations. Each eliminates fine-grained cues (kinematic cues on one hand, spectral cues on the other), both obscuring the identity of the actor/talker and greatly constricting the amount of information present in the signal. Analogously, Knoblich et al. (2002) report that self-advantages for the recognition of hand-drawn symbols are eliminated when kinematic idiosyncrasies are constrained by limiting the drawing space.

The primary goal of this experiment was to determine whether and to what extent talker intelligibility may be determined by talker-listener similarity or talker prototypicality. Talker-listener similarity was found to be a significant predictor only in SPIN and Filt-SPIN models, as may be expected given that the results also demonstrate a small advantage for self-produced stimuli in these sessions. Prototypicality, however, was found to be a significant predictor in all three models, and where present, the effects of talker-listener similarity were consistently weaker than the effect of talker prototypicality. This is consistent with previous research that found greater recognition accuracy for words produced by a statistically average speaker as compared to those produced by the participants themselves (Ch. 4)

Perhaps the most surprising result from this experiment is that prototypicality effects on word recognition hold even when a person is listening to their own recorded speech. This suggests that speech perception is strongly influenced by the stochastic history of input from a range of talkers in our linguistic communities. While it would be difficult to determine from this experiment, our results may suggest that while some speech

production processes involve comparison of an intended and a produced sound (Houde and Nagarajan, 2011), self-monitoring may also involve mechanisms and representations utilized during the perception of others (Levelt, 1983; Lind et al., 2014). While there does appear to be some role of talker-listener similarity at work it does not seem to play as *crucial* a role as talker prototypicality, casting doubt on theories that similarity-based mechanisms function as the basis of speech perception. Our results suggest that speech perception operates under the same mechanisms as other perceptual processes considered to be more "general" (Carbonell and Lotto, 2014; Holt and Lotto, 2008; Holt et al., 2010), such as face (Cabeza et al., 1999) and voice perception (Latinus and Belin, 2011; Latinus et al., 2013).

The results suggest that prototypicality and similarity mechanisms may interact during speech perception. Specifically, the significant three-way interaction found in the FiltSPIN session indicates that these two variables can modulate each other. Model estimates suggest that early in the sentence, before much input has been acquired, similarity has its strongest effects when the speaker is more prototypical. This may point to the existence of parallel emulatory processing during perception (Jeannerod, 2001; Knoblich and Flach, 2001) that can only be instantiated when the incoming speech falls close enough to the prototypical values of the relevant speech categories (Poeppel et al., 2008). Late in the sentence, however, similarity has its strongest effects when the talker is *aprototypical*, after the accumulation of talker-specific phonetic cues. This is consistent with results that suggest that listeners construct a model of the talker in order to emulate and predict how upcoming speech will sound (Brunelliere and Soto-Faraco, 2013). The fact that this interaction was only found when the recorded stimuli had been filtered to approximate how speech sounds during self-monitoring (Puria and Rosowski, 2012; Vurma, 2014) may indicate that listeners store and access representations that are based on their own sensorimotor experience as speech producers, consonant with ideomotor theory (Greenwald, 1970; Hommel et al., 2001; James, 1890).

This experiment focused solely on perception of native, same gender talkers of similar ages and linguistic backgrounds. When attempting to generalize these results to natural speech, it is important to take into account social variation. It is unlikely that all incoming speech sounds of a given category are compared to one single prototype. Listeners extract information about a talker's idiosyncratic speech style and can use this information during perception (e.g., Magnuson and Nusbaum, 2007; Remez et al., 2011). At the same time, experience with a range of talkers leads to better perception of novel talkers (Bradlow and Bent, 2008), suggesting that abstraction is also at work. This supports models of speech encoding that take into account social encoding (Johnson, 2006; Sumner et al., 2014) as well as norm-based processes.

Our findings suggest that incoming speech is processed by reference to speech prototypes that emerge from a multi-dimensional acoustic space. When a particular talker's speech patterns are closer to the prototypical values for the relevant acoustic cues, that talker's speech is more readily understood. The ability to converge on speech prototypes may explain why so many individuals with widely varying listening experience can nevertheless agree on how intelligible they find a given talker to be. The results further suggest that intelligibility may be modulated by phonetic prototypicality as well as similarity, possibly as separate mechanisms working in parallel, challenging a strict division between more embodied and more abstract theories. Rather, seemingly disparate mechanisms and representations may complement each other and work in parallel in guiding speech perception.

5.4 Appendix. Automatic Alignment Validation

Automatic forced alignment was implemented in Praat (Boersma and Weenink, 2016) utilizing Praatalign (Lubbers and Torreira, 2016). A modified version of the FormantPro toolkit was utilized for data smoothing prior to analysis (Xu, 2007). Inspection of the automatically aligned sound files indicated that, indeed, automatic alignment performed very well in most cases but still contained several clear errors. For this reason, the first author performed a manual alignment on all recordings from a single talker (Participant 9) and compared these values to those obtained automatically. This enabled us to quantify not only the accuracy of the automated alignment process overall, but also the alignment accuracy for each individual segment. To be conservative, we then only utilized those segments for which automatic alignment was found to be sufficiently accurate, and excluded those segments for which values obtained automatically and values obtained manually differed substantially.

First, we focused on differences in duration values by averaging across all instances of a given segment for the manually and automatically obtained alignments. Average segment durations for manual alignment were then subtracted average segment durations obtained via automatic forced alignment. The mean-absolute-difference between manual and automatic averages was 27ms ($max = 217ms, min = 1.4ms, sd = 37ms$). This appeared to depend on both segment quality and number of tokens. Comparison with supervised alignment suggests that as the number of tokens increases, the difference in averages obtained between automated and forced alignment techniques diminishes. The average difference in duration of segments with more than 200 occurrences was always less than 20ms: [t, d, n, r, l, s, ə]. We set this as our threshold for inclusion for all segments. Additionally, several segments with few occurrences were nevertheless

aligned with excellent accuracy ($< 10ms$), and these were also included for analysis: [ɛɪ, h, ɪ, o, u, v]. Several segments showed extreme differences between automatic and manual alignment (greater than 50ms difference), and were immediately excluded from analysis: [j, ɣ, ŋ]. This left a number of segments, [a, b, f, w, x, ʏ, z], with less than 200 occurrences that differed by less than 20ms between manual and automatic alignment. To determine whether these segments indeed met our cutoff, we validated these measurements using manually aligned measurements from another participant. We randomly chose another participant (Participant 14) and selected 50 recordings for manual alignment (also chosen at random). This data set contained three fewer segments than the full data set, specifically, [ɑu, j, n]. Comparison with automatic alignment revealed the same trend; alignment duration differences were found to be smaller when the number of tokens was greater (less than 20ms for tokens occurring more than 80 times in the sample). Of the unclassified segments, all except [w] differed between manual and automatic alignment by less than 20ms. The final set of segments included for duration analysis included [b, t, d, n, l, r, f, v, s, z, x, h, ə, ɛɪ, ɪ, o, u, a, y], totaling 19 out of the total 31 segments included in the corpus, spanning a range of manners and places of articulation.

In addition to durational measures, we also calculated, for all segments, the averages of the first 12 mel-frequency cepstral coefficients (MFCCs; Davis and Mermelstein, 1980) plus the zeroth coefficient (which corresponds to a scaled measure of energy in decibels). MFCCs measure the energy in the spectrum of a speech signal within triangular filterbanks spaced equally along the mel-scale, a psychoacoustic scale for measuring equal distances in the frequency domain which is linear below 1000 Hertz (Hz) and logarithmic above 1000 Hz (Stevens, 1937). MFCCs allow for quantification of the amount of energy in specific frequencies that distinguish particular phones, such as the high frequency energy found in the fricative sounds [s] and [ʃ].

Again, manually aligned values were compared to those obtained from automatic alignment; over all segments and coefficients, automatic alignment differed from manual alignment by an average of 0.65 standard deviations. Utilizing median instead of mean values (in order to reduce the influence of outlier values) reduced this to 0.62 standard deviations. As with duration, certain segments were found to be more amenable to automatic alignment than others, however, these were not the same segments as with the duration measurements. For example, across all MFCCs, the difference between median values for manual and automatic alignment of [n] was 1.87 standard deviations, but only 0.72 standard deviations for [ɣ]. We therefore elected to utilize median values for segments, and set a cutoff of 0.5 standard deviations across all MFCCs. This comprised 23 segments: [p, b, d, f, v, s, z, x, h, a, ɑ, ɛ, e, ɛɪ, i, ɪ, ɔ, o, øː, œy, u, ʏ].

For vocalic segments, we measured fundamental frequency (F0) and the first three resonant frequencies (formants) of the vocal tract (F1, F2, F3). Across all segments and formants, average absolute differences between manual and automatic alignment were 20.9 Hz. Again, this differed for specific segments; [ɑu] differed on average by 60.18 Hz, but [ɛ] differed by only 7.53 Hz. Small differences in segment alignment are likely to have influenced averaged results due to the inclusion of time points corresponding to flanking segments. We therefore utilized a simple algorithm to constrain measurements to their most likely values, proceeding from two assumptions: 1) That the majority of data points in a given measurement window will have been generated by the target segment, and 2) that the rate of change between sequential measurements will be greater at segment transitions than at segment midpoints. For each segment token (e.g., the first "i" in "de laars is boven de biet"), we calculated for each formant the difference between sequential samples in order to obtain the rate of change. We then excluded data points corresponding to greater or less than one standard deviation from the median of this rate of change, and recalculated averages. Applying this algorithm reduced average differences to 16.32 Hz. Further iterations failed to decrease differences between data obtained from manual and automatic alignment. We then excluded vocalic segments for which measurements between automatic and manual alignment differed by more than ± 25Hz, leaving the following segments for analysis: [ə, ɑ, a, ɛ, e, ɛɪ, i, ɔ, o, øː, œy, u, ʏ].

Finally, several additional variables were computed from the extracted formant measurements. These included the dispersion between F0 (fundamental frequency) and the first two resonant frequencies of the vocal tract (F1, F0), as F1-F0 dispersion has been shown to be an important cue for vowel height (Fahey et al., 1996; Hoemeke and Diehl, 1994; Traunmüller, 1981), while F2-F0 dispersion contributes to the perception of closed vowels (Savariaux et al., 1999). Additionally, both measures have been found to covary with F0 (Chládková et al., 2009).

In summary, for each talker we measured average sentence duration, F0, range F0, amplitude, and range amplitude (five variables). Utilizing automatic alignment, we also measured for each talker, for each vocalic and consonantal segment (for which automatic alignment was sufficiently accurate), average duration, amplitude (MFCC 0), and the average value of the first twelve MFCCs (14 variables per segment). In addition, for vocalic segments we also measured F0, the first three resonant formants (F1, F2, and F3), and F1-F0 dispersion (five variables per vocalic segment).

Chapter 6

General Discussion

With the advent of the spectrogram (Potter, 1946), phoneticians who attempted to decipher the speech code swiftly realized that there is no one-to-one correspondence between articulation and acoustics (Shankweiler and Fowler, 2015). For decades, researchers have investigated how listeners solve the problems of *segmentation*, extracting discrete units from a continuously varying signal, and *invariance*, recognizing distinct acoustic signatures as a single unit. To solve the problems of segmentation and, in particular, invariance, researchers in speech perception have often characterized the targets of speech perception in articulatory terms (Best, 1995; Fowler, 1986; Galantucci et al., 2006; Liberman and Mattingly, 1985, 1989). Meanwhile, researchers in speech motor control have consistently characterized the targets of speech production as primarily acoustic (e.g., Houde and Nagarajan, 2011; Niziolek et al., 2013). Rather than attempt to prove whether representations for speech are ultimately acoustic or articulatory in nature, this thesis set out to examine a simpler question: Does the experience of producing speech and hearing the consequences of our own productions, i.e., sensorimotor experience, influence how we perceive the speech of ourselves and others?

In attempting to answer this question, the studies presented in this thesis often involved novel experimental paradigms. Thus, many of the results would fall under the category of exploratory rather than confirmatory research (Wagenmakers et al., 2012). That being said, we observed patterns of results across all four experimental chapters that clarified the role of sensorimotor experience in speech perception.

Chapter 1 provided a short overview of sensorimotor control theory for speech production, and suggested several paths by which sensorimotor experience may affect perception. In Chapters 2 and 3, we examined whether exposing participants to altered auditory feedback, thus introducing a discrepancy between production and perception, led to changes in the categorization of speech sounds. Our results suggested that when acoustics and articulation diverged, behavior in subsequent perception tasks followed the direction of articulation. In Chapters 4 and 5, we examined how long-term sensorimotor experience may influence the recognition of words produced by oneself as well as by other speakers. In contrast to Chapters 2 and 3, our results suggested that sensory, not sensorimotor experience, was a better predictor of word recognition.

In the following sections, I first discuss the implications of Chapters 2 and 3 and Chapters 4 and 5 separately. I then discuss how these results can be reconciled within recent neurobiological models of speech production and perception (Hickok and Poeppel, 2004, 2007), and suggest candidate roles for sensorimotor experience in speech perception.

6.1 Mapping acoustics to articulation in phonetic categorization

Chapter 2 investigated the role of sensorimotor experience in the perception of coarticulated speech sounds. In two experimental sessions separated by at least two weeks, native speakers of English performed interleaved production and perception tasks. Interleaving these tasks enabled us to examine whether the categorization of fricatives coarticulated with a following vowel (Kunisaki and Fujisaki, 1977) was affected by recent sensorimotor experience. In the first session, auditory feedback was unaltered. In the second session, the auditory feedback that participants heard during production of words containing the high front vowel [i] was shifted. By lowering the value of the second resonant frequency of the vocal tract (F2), participants heard themselves producing a more [u]-like vowel than intended. Previous experiments (Lametti et al., 2014b; Shiller et al., 2009) have found that most participants exhibited opposing articulatory responses to such shifts in feedback. Such opposing responses counteract the shift in acoustics. Accordingly, many of our participants also opposed the shift in auditory feedback by hyper-articulating the vowel [i], while a few failed to exhibit any changes in articulatory behavior.

Interestingly, and not entirely unexpectedly (MacDonald et al., 2011), some participants exhibited significant changes in production *following* the direction of the shifted feedback. That is, in response to hearing themselves producing [i] with a lower F2 than normal, these participants articulated [i] with a lower F2 than normal. Contrastively, participants that opposed the shifted feedback heard a vowel with a lower F2, yet articulated a vowel with a higher F2. By comparing participants with divergent behavioral responses to the production task, we were able to directly test whether changes in the categorization task reflected auditory exposure (i.e., what participants *heard* during the production tasks) or articulatory behavior (i.e., what they *produced*).

Comparing across the two sessions, changes in the perceptual task were found to correlate with what participants *articulated* rather than what they heard. We interpreted these results as evidence of a "remapping" between articulation and acoustics. Adaptation to the altered feedback (whether opposing or following) altered the relationship between

a given articulatory motor program and the resulting acoustic consequences. This newly acquired mapping was then carried over into the perceptual task, resulting in a shift in the perceptual boundary of the coarticulated fricative. What is particularly interesting about this finding is the fact that the stimuli for the production task consisted of only stop consonants and vowels. No fricatives were uttered during the production task, ruling out the possibility that the change in perception could be attributed to response bias or sensory exposure.

The results of Chapter 2 indicated that sensorimotor experience may update mappings between articulation and acoustics during production and that these mappings are utilized when categorizing speech sounds. Chapter 3 expanded upon these results by examining how sensorimotor mappings may be reflected in electrophysiological responses to a phonetic continuum. In a purely perceptual experiment utilizing electroencephalography (EEG), Bidelman et al. (2013) found that the perception of a phonetic continuum varying along F1 between [u] and [a] (i.e., varying in the openness/closedness of the jaw and tongue body) elicited distinct cortical responses depending on the F1 value of the stimulus step. Specifically, the amplitude of the P2 component was found to correspond to participants' judgments of vowel identity, even for different classifications of the same acoustic stimulus. Based on these results, we hypothesized that changes in P2 amplitude may reflect sensorimotor remapping, as observed in Chapter 2.

In the experiment described in Chapter 3, native speakers of Dutch repeatedly produced the Dutch word 'pet' (English: "cap"), containing the vowel [ɛ]. These participants then classified stimuli along a five-step continuum between [ɛ] and [ɪ]. The same participants then produced the word 'pet' again. However, in this second speaking task, auditory feedback was shifted, causing participants to hear themselves producing a vowel more like the English [æ] in "pat". Twenty out of twenty-eight participants exhibited significant opposing responses, articulating a vowel more like the [ɪ] in "pit". These "adapters" then performed the same phonetic categorization task. We compared the results of the twenty adapters to twenty control participants for whom feedback was unaltered.

In contrast to a similar experiment (Lametti et al., 2014b), we did not observe significant differences in perceptual changes between the two groups. However, when we examined individual behavior profiles, we found evidence for interactions between changes in articulation and changes in perception. Furthermore, we found that individual differences regarding produced F1 and F2 were reflected in cortical responses. By comparing our electrophysiological results with those found in a previous sensorimotor adaptation experiment (Ito et al., 2016), we were able to conclude that the observed changes in P2 amplitude reflected a remapping between articulation and acoustics.

The experimental method employed in Chapters 2 and 3 enabled us to hold the perceptual stimulus constant but alter the relationship between articulation and acoustics. The primary finding that emerged from these two chapters was that phonetic categorization appeared to involve a transformation of acoustic information onto articulatory representations. In Chapter 2, when auditory feedback was manipulated such that acoustics and articulation diverged, the subsequent changes in perception followed articulation rather than acoustics (page 28, Fig. 2.3). Similarly, in Chapter 3, we found that changes in perception correlated with changes in production (page 49, Fig. 3.3C). Furthermore, for participants who received altered auditory feedback, fluctuations in the average amplitude of the event related P2 component during perception correlated with articulatory rather than acoustic measures (page 53, Fig. 3.5C). A purely auditory account (Diehl et al., 2004; Lotto and Kluender, 1998) is unable to account for this pattern of observed results.

Our results suggest that phonetic categorization involves a sensorimotor transform that draws on the listener's recent experiences during speech production. Altering the relationship between articulation and acoustics during production tasks led to subsequent changes in behavioral and neural responses in perception tasks. The original motor theory suggested that associations between articulation and acoustics are acquired during development (Liberman et al., 1957). In this regard, it had much in common with traditional ideomotor theories (James, 1890) that posited that links between action and perception are created by repeated experience of stimulus - action - feedback chains (Greenwald, 1970). While later versions of the motor theory abandoned the ontogenetic relationship between articulation and acoustics in favor of a phylogenetic account (Liberman et al., 1967), our findings suggest that speech sound perception may indeed be mediated by ongoing sensorimotor experience.

6.1.1 Sensorimotor experience alters probabilistic sound categorization

One way to characterize the mediating influence of sensorimotor experience, while remaining neutral on the question of whether representations are "ultimately" motoric or acoustic, is by recourse to probabilistic models of speech perception (Clayards et al., 2008; Kleinschmidt and Jaeger, 2015b; Norris and McQueen, 2008). This approach posits that speech perception is a problem of inference under uncertainty and has been successful in modeling the effects of selective adaptation and phonetic recalibration (Kleinschmidt and Jaeger, 2015a), the perceptual magnet effect (Feldman et al., 2009), and compensation for coarticulation (Sonderegger and Yu, 2010). In phonetic categorization, the task of the listener is to decide which category C some acoustic stimulus S belongs to. Yet due to variability in speech motor control (Houde and Nagarajan, 2011),

speakers do not always hit their intended speech target (Niziolek et al., 2013). Listeners assume for this reason that *S* is normally distributed around a target pronunciation *T*, which itself is distributed around a category mean (i.e., an articulatory motor program given an acoustic target; Sonderegger and Yu, 2010). Thus characterized, the task of phonetic categorization fits well within an articulation-based representational framework (Halle and Stevens, 1962; Liberman and Mattingly, 1985; Poeppel et al., 2008), as the goal of the task is to infer the speaker's intended motor program from its actual realization. Through sensorimotor adaptation, the likelihood that S_i was generated by T_i shifts, such that the intention to produce a speech sound category C_i requires implementing a different articulatory gesture T_i'. By changing the values of pronunciation *T*, sensorimotor experience thus changes the probability of category *C* having generated the acoustic value *S*.

The finding that altered perceptual experience can shift a perceptual boundary is not novel. Phonetic recalibration can be triggered by biasing lexical (Norris et al., 2003; van Linden et al., 2007) or visual (Bertelson et al., 2003) information that leads listeners to update their beliefs about the relationship between an acoustic stimulus and a phonetic category (Kleinschmidt and Jaeger, 2015b). What is interesting about sensorimotor adaptation, however, is the pattern of *generalization*. Phonetic recalibration is notoriously talker specific (Eisner and McQueen, 2005; Kraljic and Samuel, 2007, though see Kraljic and Samuel 2006), and does not occur when variation in pronunciation can be attributed to an incidental source (e.g., a pen in the speaker's mouth; Kraljic et al., 2008). In Chapters 2 and 3, we found evidence suggesting that participants generalized from their own sensorimotor experience to the perception of the speech of others. This suggested that while sensory adaptation may rely on perceptual similarity between the adapted sound and the test sound, sensorimotor adaptation targets a different level of probabilistic inference. However, the results of Chapter 3 showed clear effects for the unambiguous training vowel and the most ambiguous stimulus step, but not the remaining stimulus steps, suggesting some degree of specificity to sensorimotor adaptation effects as well.

To summarize, the results of Chapters 2 and 3 suggest that phonetic categorization can be biased by recent sensorimotor experience and that perceptual changes stemming from sensorimotor experience can be characterized probabilistically in terms of an updating of beliefs. Furthermore, the results demonstrate that we generalize our own sensorimotor experience to the perception of others. However, as stated in Chapter 1, tasks involving explicit phoneme identity judgments appear to recruit different processing resources than other speech perception tasks (Hickok and Poeppel, 2000, 2004, 2007; Stasenko et al., 2015). Therefore, in Chapters 4 and 5, we examined the possible role of sensorimotor experience using word recognition tasks.

6.2 Sensory vs. sensorimotor experience in word recognition

In Chapters 4 and 5, we examined to what extent sensorimotor experience may play a role in the identification of degraded words. In order to do so, we operationalized sensorimotor experience in terms of the acoustic similarity between talker and listener. If listeners map incoming speech onto representations formed by their own sensorimotor experience, then perception should be facilitated to the degree that the perceived word matches how listeners would produce that same word themselves. Furthermore, similarity should reach its maximal value for self-produced stimuli, as listening to one's own recordings would constitute the greatest possible match to one's own production experience (Tye-Murray et al., 2013, 2015).

Chapter 4 tested the intelligibility of words produced by an average speaker compared to self-produced words. In this experiment, we recorded twenty-eight female participants producing single Dutch nouns. These words varied in production frequency (i.e., how often a word occurs in conversation or print) and phonological neighborhood density (i.e., how many words exist in Dutch that differ from this word by a single segment). We then presented these participants with six-band, noise-vocoded versions of these words (Shannon et al., 1995). This manipulation eliminated fine-grained spectral cues to talker identity (López et al., 2013) while preserving individual variation in durational cues, as well as differences in the average amplitude of the six frequency bands.

After all recordings had been collected, we conducted a phonetic analysis of the recordings in order to identify a "statistically average" speaker. Participants then listened to random stimuli drawn from their own recordings as well as the recordings of the average speaker and attempted to identify the degraded words. The intelligibility of the average speaker thus served as a baseline against which we could measure facilitation for self-produced speech. However, in contrast to our predictions, participants were more accurate at identifying words produced by this average speaker compared to their own recordings.

In a follow-up experiment, we recruited a new set of participants who performed the same word identification task. However, each participant was presented with a subset of words from all twenty-eight speakers who participated in the first experiment. By averaging accuracy for each talker across listeners, we obtained general intelligibility scores for the participants in Experiment 1. Two interesting results were found: First, we found that if a talker had a lower general intelligibility score, this talker was also more likely to have been less accurate at identifying words when listening to self-produced stimuli. This suggested that listening to recordings of one's own speech is much like listening to another talker. Second, we found that the "average-speaker" advantage

remained significant even when taking into account differences in general intelligibility. The results of the second experiment thus confirmed that the increased accuracy for the average speaker compared to one's own voice could be attributed to the statistical averageness of the speaker and was not due to simply having accidentally selected the most intelligible speaker of the sample.

The results of Chapter 4 suggested that a talker is more intelligible when their speech patterns align more closely to the statistical average of their linguistic community. Yet it is difficult to generalize from noise-vocoded speech to natural conversation. In natural conversation, speakers *do* have access to fine-grained spectral cues and often know who it is that they are talking to. For this reason, we set out to perform a more comprehensive investigation of the role of statistical averageness compared to similarity, and in so doing, compare the effects of sensory vs. sensorimotor experience on speech perception.

Chapter 5 constituted a systematic investigation of the effects of talker prototypicality compared to talker-listener similarity. As in the previous study, we operationalized sensorimotor experience in terms of acoustic similarity of a talker's speech patterns to those of the listener. Talker prototypicality was operationalized as the average similarity of one talker's speech patterns to all other same-gender talkers in the sample. Instead of listening to isolated words, participants attempted to identify two target words in a simple sentential context. In addition to listening to their own recorded stimuli, participants were presented with stimuli from six other randomly selected talkers. Furthermore, for half of the participants, each experimental trial was accompanied by a label indicating the identity of the talker, ensuring that these participants could track different talkers and were aware of when they were listening to their own recordings.

When presented with noise-vocoded stimuli, we found that listeners were no better at recognizing self-produced words than words produced by other talkers. Furthermore, we found that the prototypicality of a talker, but not talker-listener similarity, was a significant predictor of whether or not a word would be accurately recognized. No effect of an accompanying label telling the participants whether they were listening to themselves or another person was found. Thus, when fine-grained spectral cues were removed by noise-vocoding, listeners showed no advantages for self-produced speech or speech that was similar to their own sensorimotor experience (as in Chapter 4). However, listeners did show advantages for prototypical speech, indicating that the effects of statistical averageness survived spectral degradation.

When presented with words embedded in speech-shaped noise, which in contrast to noise-vocoded words retained fine-grained spectral information, listeners did indeed show an advantage for recognition of self-produced words compared to the average of

six random talkers. Accordingly, talker-listener similarity was found to be a significant predictor of accuracy. However, the magnitude of the effect of prototypicality was much larger than the effect of similarity. Again, there was no effect of an accompanying label. Taken together, these results suggested that sensorimotor experience facilitated word recognition, yet this effect was much weaker than the effect of prototypicality.

The role of sensorimotor experience in speech perception was further clarified by a third manipulation, in which listeners were presented with stimuli embedded in speech-shaped noise that had first been filtered to approximate the effects of bone-and-air conduction (Békésy, 1949). These caused the stimuli to sound more like how we hear ourselves "in our heads" when we speak (Vurma, 2014). As found with the unfiltered stimuli, listeners were more accurate for self-produced speech compared to the average accuracy for the speech of six other talkers. Additionally, models containing prototypicality and similarity as predictors revealed a three-way interaction between how prototypical a talker's speech was, how similar the talker was to the listener, and the position of the target word in the sentence. When the target word appeared early in the sentence, similarity had a strong effect when the talker was prototypical but almost no effect when the talker was aprototypical. This effect was reversed when the target word appeared late in the sentence; similarity had a small effect when the talker was prototypical and a stronger effect when the talker was aprototypical (page 106, Fig. 5.3). Even so, prototypicality was a much stronger predictor of accuracy than similarity, pointing to the primacy of sensory over sensorimotor experience in word recognition.

6.2.1 Word recognition draws on representations abstracted over multiple talkers

Chapter 5 is related to other studies that have examined the role of acoustic variables in determining speech intelligibility (e.g., Bradlow et al., 1996; Hazan and Markham, 2004; Hirsh et al., 1954). However, rather than examining individual acoustic variables such as average pitch or vowel dispersion (Bradlow et al., 1996), we extracted statistical information across multiple acoustic dimensions (Holt et al., 2010). Our results demonstrated that the prototypicality of a talker's speech with respect to the speech community at large was a stronger predictor of intelligibility than the similarity of that talker's speech to their own speech. This effect of prototypicality holds even when we are listening to recordings of ourselves. If our own speech patterns are further from the statistical average of the linguistic community, we are also more likely to find our own speech less intelligible (even when we know that we are listening to ourselves).

Evidence from second language acquisition suggests that the relative contribution of sensory compared to sensorimotor experience may depend on the amount of experience one has had listening to a range of talkers in a specific language (e.g., Baese-Berk et al., 2013; Bradlow and Bent, 2008; Lively et al., 1993). When learning the English /l/ vs. /r/ distinction, Japanese learners tend to make fewer misidentifications of self-produced words than words produced by native speakers of English (Sheldon and Strange, 1982). While the sample size of this study was limited (six participants), a similar pattern of results was also found in a small group of Korean participants (Borden et al., 1983). The findings of Chapters 4 and 5 suggest that the self-perception advantage found for second-language learners may reflect the fact that not enough sensory input has accumulated in order to generate native-like auditory prototypes.

The results of Chapters 4 and 5 suggest that sensory input from a wide variety of talkers leads listeners to converge on perceptions of what constitutes prototypical speech in their community. Such convergence is quite surprising given that, over our lives, we regularly interact only with a small subset of the population (Hill and Dunbar, 2003). Furthermore, of that small subset, those who we interact with face-to-face for extended periods of time comprise an even smaller subset of that input (Lev-Ari, 2015). Accordingly, researchers have cautioned their readers to limit interpretations of intelligibility measurements to the specific "crew" of talkers and listeners tested (Bradlow et al., 1996; Hirsh et al., 1954). This is excellent practice, as it does not proceed from the assumption that listeners' perceptions are necessarily similar, and in research and clinical settings it remains extremely important to consider possible influences of the speaker, listener, and task (McHenry, 2011; Miller, 2013). That being said, listeners have been found to exhibit high inter-rater agreement when judging accentedness, fluency (Pinget et al., 2014) as well as intelligibility (Doyle, 1987; Hazan and Markham, 2004). Based on our results, such high agreement in ratings may be reflective of a general auditory process (Holt et al., 2010) that encodes information in terms of prototypes (Kuhl, 1991; Rosch, 1973).

6.3 Dual—streams for phonetic categorization and word recognition

Chapters 2 and 3 suggest that sensorimotor experience modulates the categorization of speech sounds, while Chapters 4 and 5 suggest that sensory input is more important for word recognition than sensorimotor experience. As mentioned earlier, these results appear to be in conflict, yet only if one assumes that the two perceptual tasks employed,

phoneme identification and word recognition, are both implemented by the same processing system. However, as has been argued in Chapter 1, ample evidence suggests that tasks that differentially emphasize lexical or sub-lexical units may recruit different neural processing resources (Hickok and Poeppel, 2000). Dissociations in speech tasks as well as evidence from clinical populations has led to the development of "dual-stream" models of speech perception (Hickok, 2012a; Hickok and Poeppel, 2004, 2007; Norris et al., 2000; Rauschecker and Scott, 2009).

Like their counterpart in vision science (Goodale and Milner, 1992), dual-stream models suggest that there are two broad pathways utilized for perceptual processing. In speech perception, the ventral pathway is proposed to map acoustic input onto conceptual-semantic representations, while the dorsal pathway is responsible for sound localization (Rauschecker, 1998) and sensory-motor integration (Hickok, 2012b; Hickok and Poeppel, 2004). For this reason, these two streams have been broadly described as the "what" and "where/how" pathways (Belin and Zatorre, 2000). Accordingly, neuroimaging of overt repetition of pseudowords (vs. real words) and attentive listening to meaningful speech (vs. meaningless speech) reveals fiber tracts corresponding to the dorsal and ventral streams (Saur et al., 2008). The implication is that lexical processing predominantly activates the ventral stream to enable speech identification, while sensorimotor integration activates the dorsal stream to map acoustic information onto articulatory representations, enabling speech production (Doupe and Kuhl, 1999; Guenther, 2006; Tourville and Guenther, 2011).

Within this framework, it is likely that the effects observed in Chapters 2 and 3 predominantly reflect activity in dorsal stream processes, while the effects observed in Chapters 4 and 5 predominantly engaged the ventral stream. The results of Chapters 4 and 5 suggest that sensorimotor experience is not as important as sensory experience in word recognition. This accords with accounts that have proposed a 'modulatory' role for sensorimotor systems in speech perception (Hickok, 2012b; Hickok et al., 2011) and have emphasized the role of the ventral stream in phoneme and word recognition (DeWitt and Rauschecker, 2012; Scott et al., 2000). As evidence, researchers often point to clinical studies suggesting that auditory word recognition is spared when there is damage to speech production systems, as in the case of Broca's aphasia (Moineau et al., 2005). Furthermore, evidence suggests that performance in phoneme identification tasks dissociates from performance in "ordinary" speech recognition tasks (Carbonell and Lotto, 2014; Hickok, 2009; Hickok and Poeppel, 2000).

Though not as strong as prototypicality, talker-listener similarity was found to play a role in predicting word intelligibility. This may suggest that conscious speech perception involves integrating the output of the dorsal stream, which maps incoming speech onto

sensorimotor experience, and the ventral stream, which maps incoming speech onto sensory representations. If these parallel streams operate in isolation, then the ultimate percept may involve an integration of the outputs of the two streams, similar to multimodal integration (Erickson et al., 2014; Marques et al., 2014; McGurk and MacDonald, 1976).

Alternatively, the results of Chapter 5 suggest that the implementation of sensorimotor processes may in some way depend on the prototypicality of the input. In Chapter 1, we suggested that sensorimotor experiences may reflect emulatory processing (Grush, 2004; Pickering and Garrod, 2007; Tian and Poeppel, 2013). However, in order for an emulatory process to be instantiated, the listener needs to already have an idea of *what* to emulate. Accordingly, Poeppel et al. (2008) suggest that listeners construct a fast auditory sketch of incoming input based on salient acoustic features (Stevens, 2002). It may be that prototypical speech that best matches an acoustic representation instantiates a sensory-to-motor transform of the input, which can enhance perception when there is a greater match between the signal and the sensorimotor representation (e.g., page 106, Fig. 5.3, left panel). This would lead us to reinterpret the lack of similarity effects for noise-vocoded speech as either the signal lacking sufficient typical cues that instantiate these processes, or that the emulated signal is an insufficient match to the spectrally averaged speech to facilitate perception.

Situating our results within biologically-motivated models of speech production may shed light on debates about the nature of the representations for speech. Both direct realism (Best, 1995; Fowler, 1986) and the motor theory (Galantucci et al., 2006; Liberman and Mattingly, 1985, 1989) posit that the ultimate objects of speech perception are not the proximal auditory signal but the distal articulatory gesture that produced the target speech. The two theories differ in that direct realism suggests that we perceive the *actual* articulatory gesture, at a sub-phonemic level, while motor theory posits that listeners recover the *intended* gesture, undistorted by coarticulatory influences. Conversely, general auditory accounts posit that the percepts of speech perception are acoustic representations (Diehl et al., 2004; Holt et al., 2010), while common coding theories argue for shared representations stored in a common representational medium (Guenther, 2006; Hommel et al., 2001; Prinz, 1990). This assumes that there exists a single type of representation utilized in speech perception. However, speech motor control theory suggests that speech production involves acoustic, somatosensory and proprioceptive representations (Houde and Nagarajan, 2011; Tourville and Guenther, 2011), which may be active to different degrees depending on the situation (e.g., auditory representations may not be as important as somatosensory representations when feedback is masked).

What effects from various speech perception tasks (Hickok and Poeppel, 2000, 2004) and the results of this thesis suggest is that there is no singular type of representation utilized for speech perception either (Evans and Davis, 2015). Rather, the goals of the perceptual task (or individual differences, e.g., Nuttall et al., 2016) may emphasize different representations to differing degrees. Given evidence that different perceptual tasks appear to activate dorsal or ventral streams to different degrees (Saur et al., 2008), apparent conflicts regarding the nature of the representations for speech perception may be resolved by reference to distinct neural processing streams.

6.3.1 Candidate roles for sensorimotor experience in speech perception

While it is clear that sensorimotor experience is not crucial for understanding clearly spoken words in one's native language and accent, sensorimotor experience may play a role in other contexts. For example, when speech is low-pass filtered and compressed in time, listeners with Broca's aphasia perform much more poorly than controls (Moineau et al., 2005). Similarly, in an experiment with healthy listeners, Nuttall et al. (2016) recently found that accurate recognition of distorted syllables correlated with activity in motor areas. These and similar findings have led many researchers to suggest that sensorimotor experience and sensorimotor processes may be important for the perception of degraded or distorted speech (D'Ausilio et al., 2012; Davis and Johnsrude, 2007; Nuttall et al., 2016; Scott et al., 2009).

The results of this thesis suggest that, in addition to degradation or distortion, sensorimotor experience may be important when dealing with atypical speech patterns, such as in the case of accented speech. For example, listeners with various types of aphasia have also been found to make significantly more errors than controls in the comprehension of accented speech (Dunton et al., 2010), suggesting that unfamiliar accents exacerbate comprehension difficulties. Similarly, healthy listeners have been found to be slower at processing unfamiliar accents (Adank et al., 2009; Adank and McQueen, 2007).

The processing of distorted and accented speech recruits ventral and prefrontal areas of the dorsal stream (Adank et al., 2015), pointing to the involvement of extrasensory processing areas for such speech. In a direct demonstration that sensorimotor experience can increase the intelligibility of accented speech, Adank et al. (2010) found that imitation of an artificial foreign accent improved comprehension more than simple auditory exposure. Similarly, utilizing a battery of training conditions, Borrie and Schäfer (2015) found that overt imitation (paired with written feedback) led to the greatest increases in the intelligibility of dysarthric speech. These results suggest that mapping acoustics to articulation through production can improve the comprehension of atypical

speech (though evidence also suggests that accent learning can be mitigated by native accent phonology; Nguyen et al., 2012). In the absence of overt production, sensorimotor processes may be instantiated through mental imagery of expected sounds (Tian and Poeppel, 2010, 2013). This may serve an online function that helps model atypical speech (Moulin-Frier and Arbib, 2013; Moulin-Frier et al., 2012). Evidence from neuroimaging also suggests that active imitation may increase recruitment of sensorimotor integration processes during subsequent perception, which may aid perceptual learning of a foreign accent (Adank et al., 2013). This is in accordance with other findings (Nuttall et al., 2016) indicating that increased involvement of sensorimotor processing, ostensibly localized in the dorsal stream, is associated with perceptual gains.

6.4 Limitations and future research

Our results suggest that sensorimotor experience may update mappings between acoustics and linguistic categories (Chapters 2 and 3), and that sensorimotor experience, though not essential, may modulate word recognition under specific conditions (Chapters 3 and 4). This may reflect the fact that speech perception is subserved by dual processing streams (Hickok and Poeppel, 2004, 2007; Rauschecker and Scott, 2009) that perform different functional roles. When generalizing our results to speech perception as a whole, the role of sensorimotor experience appears to be modulatory. Furthermore, our research into the role of sensorimotor processing has revealed that speech prototypicality appears to play an important role in auditory word recognition.

One clear finding that emerged from this thesis is that the effects of sensorimotor experience varied greatly depending on the materials, tasks, and participants involved. Therefore, in addition to considering the findings of this work, it is important to consider its limitations. A picture of speech perception based solely on two-alternative forced choice phoneme categorization tasks would be a gross mischaracterization (Holt et al., 2010). Chapters 2 and 3 support other studies that have found that altered sensorimotor experience can affect phonetic categorization (Lametti et al., 2014b; Shiller et al., 2009), yet few experiments have examined sensorimotor adaptation in the context of other factors intrinsic to speech, such as lexical variation (Bourguignon et al., 2016). Full understanding of how sensorimotor adaptation affects speech perception will require utilizing a range of manipulations, such as perturbing natural running speech rather than isolated productions of a single sound (Cai et al., 2011), as well as a variety of tasks that draw attention not only to phonetic contrasts but also semantics (Krieger-Redwood et al., 2013). Synthesizing results from diverse tasks will provide a broader

and more ecologically valid perspective on the role of sensorimotor experience in speech perception.

Chapters 4 and 5 suggest that talker prototypicality is an important predictor of talker intelligibility. Yet the experiments involved only native speakers of Dutch who listened to other same-gender native speakers of Dutch. Speech perception is known to be influenced by talker familiarity (Nygaard and Pisoni, 1998), suggesting that we may complement these abstract representations with more specific ones as well. If listeners abstract representations over groups of talkers, then the perception of a particular talker may depend on the social group to which that talker belongs (Drager, 2010; Johnson, 2006; McGowan, 2015; Sumner, 2015; Sumner et al., 2014). Listeners' expectations about how different groups should sound has been found to modulate how they perceive them to sound (Babel and McGuire, 2013; Niedzielski, 1999), as well as how listeners map acoustic information onto lexical representations (Dahan et al., 2008; Van Berkum et al., 2008). This suggests that, in certain cases, listeners' expectations may not depend on prototypicality, reflecting the actual statistical variation in the environment, but rather stereotypicality, reflecting listeners' cultural assumptions of what people (of specific social groups) are expected to sound like (Babel and McGuire, 2015). Regarding the findings of Chapter 5, this may predict that prototypicality would be a weaker predictor of intelligibility when an individual holds implicit cultural assumptions about how a given social group should sound (Babel and Russell, 2015; Rubin and Smith, 1990; Yi et al., 2013). In light of these concerns and the growing recognition of the importance of social factors in speech perception (e.g., Creel and Bregman, 2011), it is crucial to examine how sociophonetic factors may impact our results before generalizing these findings to speech perception as a whole.

6.5 Conclusion

We can learn much about speech production and speech perception by examining each separately (McQueen, 2005). Technological advancement, such as the invention of the spectrogram and the Pattern Playback device, spurred rapid advancement in our ability to manipulate and understand the acoustic speech signal, and in turn the nature of speech perception (Shankweiler and Fowler, 2015). Similarly, recent technological advances (Cai et al., 2008; Houde and Jordan, 1998, 2002; Tourville et al., 2013) offer us the ability to manipulate production and examine its effects on perception (e.g., Adank et al., 2010; Ito et al., 2016; Lametti et al., 2014b; Shiller et al., 2009) as well as manipulate perception and examine consequent effects on production (Bourguignon et al., 2016; Franken et al., 2015; Lametti et al., 2014a). If we hope to understand the

relationship between perception and production, it would be beneficial to conduct experiments that manipulate properties of one modality and observe whether this has any effect on the other, as this thesis has done.

The voice we experience most consistently in the world is our own. We hear our own voice in all sorts of acoustic environments with different reverberations, under the influence of different emotions, and in a variety of formal and informal settings. For this reason, it is understandably tempting to posit a strong role for this experience in shaping our perceptions. Yet we should only extend our estimations of these effects as far as the data permit, and the data suggest that the role of sensorimotor experience may better be described as modulatory rather than crucial. As discussed in the introduction, the purpose of having senses is to enable us to successfully navigate our environment. It is therefore crucial that perception in general, and speech perception in particular, should be sensitive to the statistical properties of sensory input that are relevant for ecologically beneficial behavior (Corney and Lotto, 2007). Our perceptual systems must find a way to deal with the problem that no two people will say a word in the same way, and not even a single person will say the same word the same way twice. This thesis suggests that, to overcome this problem, listeners may abstract over auditory input from a range of different speakers and, under certain conditions, draw on their own experience as speech producers in order to turn noisy, ambiguous, and highly variable acoustic input into meaningful language.

Bibliography

Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 35(2):520–529. 10.1037/a0013552.

Adank, P., Hagoort, P., and Bekkering, H. (2010). Imitation Improves Language Comprehension. *Psychological Science*, 21(12):1903–1909. 10.1177/0956797610389192.

Adank, P. and McQueen, J. (2007). The effect of an unfamiliar regional accent on spoken word comprehension. *In: Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS). International Congress of Phonetic Sciences: Saarbrücken, Germany. (2007)*.

Adank, P., Nuttall, H. E., Banks, B., and Kennedy-Higgins, D. (2015). Neural bases of accented speech perception. *Frontiers in Human Neuroscience*, 9. 10.3389/fnhum.2015.00558.

Adank, P., Rueschemeyer, S., and Bekkering, H. (2013). The role of accent imitation in sensorimotor integration during processing of intelligible speech. *Frontiers in Human Neuroscience*, 7.

Adank, P., Van Hout, R., and Smits, R. (2004). An acoustic description of the vowels of northern and southern standard dutch. *The Journal of the Acoustical Society of America*, 116(3):1729–1738.

Alain, C., Campeanu, S., and Tremblay, K. (2010). Changes in Sensory Evoked Responses Coincide with Rapid Improvement in Speech Identification Performance. *Journal of Cognitive Neuroscience*, 22(2):392–403.

Allen, J. S. and Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, 115(6):3171.

Babel, M. and Bulatov, D. (2011). The Role of Fundamental Frequency in Phonetic Accommodation. *Language and Speech*, 55(2):231–248.

Babel, M. and McGuire, G. (2013). Listener expectations and gender bias in nonsibilant fricative perception. *Phonetica*, 70(1-2):117–151.

Babel, M. and McGuire, G. (2015). Perceptual fluency and judgments of vocal aesthetics and stereotypicality. *Cognitive Science*, 39(4):766–787.

Babel, M. and Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, 137(5):2823–2833.

Baese-Berk, M. M., Bradlow, A. R., and Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, 133(3):EL174–80. 10.1121/1.4789864.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278. 10.1016/j.jml.2012.11.001.

Bartoli, E., D'Ausilio, A., Berry, J., Badino, L., Bever, T., and Fadiga, L. (2015). Listener-speaker perceived distance predicts the degree of motor contribution to speech perception. *Cerebral Cortex*, 25(2):281–288.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48. 10.18637/jss.v067.i01.

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4.

Beijering, K., Gooskens, C., and Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. *Linguistics in the Netherlands*, 25(1):13–24. 10.1075/avt.25.05bei.

Békésy, G. v. (1949). The structure of the middle ear and the hearing of one's own voice by bone conduction. *The Journal of the Acoustical Society of America*, 21(3):217.

Belin, P. and Zatorre, R. J. (2000). 'What', 'where' and 'how' in auditory cortex. *Nature Neuroscience*, 3(10):965–6. 10.1038/79890.

Belin, P. and Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport*, 14(16):2105–2109.

Bell-Berti, F. and Krakow, R. A. (1991). Anticipatory velar lowering: a coproduction account. *The Journal of the Acoustical Society of America*, 90(1):112–23.

Bertelson, P., Vroomen, J. J., and De Gelder, B. (2003). Visual recalibration of auditory speech identification: A mcgurk aftereffect. *Psychological Science*, 14(6):592–597.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In *Speech perception and linguistic experience: Issues in cross-language research*, pages 171–204.

Bidelman, G. M., Moreno, S., and Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *NeuroImage*, 79:201–12.

Blumstein, S. E. (1994). Impairments of speech production and speech perception in aphasia. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 346(1315):29–36. 10.1098/rstb.1994.0125.

Blumstein, S. E. and Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, 10(1-3):25–32.

Boersma, P. and Weenink, D. (2013). Praat: doing Phonetics by Computer [Computer program]. Version 5.0.1. Retrieved from http://www.praat.org/.

Boersma, P. and Weenink, D. (2016). Praat: doing Phonetics by Computer [Computer program]. Version 6.0.21. Retrieved from http://www.praat.org/.

Borden, G., Gerber, A., and Milsark, G. (1983). Production and perception of the /r/-/l/ contrast in Korean adults learning English. *Language Learning*, 33(4):499–526. 10.1111/j.1467-1770.1983.tb00946.x.

Borrie, S. A. and Schäfer, M. C. (2015). The role of somatosensory information in speech perception: Imitation improves recognition of disordered speech. *Journal of Speech, Language, and Hearing Research*, 58(6):1708–1716.

Bourguignon, N. J., Baum, S. R., and Shiller, D. M. (2016). Please say what this word is- Vowel-extrinsic normalization in the sensorimotor control of speech. *Journal of Experimental Psychology: Human Perception and Performance*, 42(7):1039–1047. 10.1037/xhp0000209.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: long-term retention of learning in perception and production. *Perception & Psychophysics*, 61:977–985. 10.3758/BF03206911.

Bradlow, A. R. and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729. 10.1016/j.cognition.2007.04.005.

Bradlow, A. R. and Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, 106(4):2074–2085.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, 101(4):2299–310.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3):255–272. 10.1016/S0167-6393(96)00063-5.

Brainard, M. S. and Doupe, A. J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. *Nature Reviews Neuroscience*, 1(1):31–40.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brown, M. and Kuperberg, G. R. (2015). A hierarchical generative framework of language processing: Linking language perception, interpretation, and production abnormalities in schizophrenia. *Frontiers in Human Neuroscience*, 9. 10.3389/fnhum.2015.00643.

Brown, R. and Berko, J. (1960). Word association and the acquisition of grammar. *Child Development*, 31(1):pp. 1–14.

Bruderer, A. G., Danielson, D. K., Kandhadai, P., and Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44).

Brunellière, A., Dufour, S., Nguyen, N., and Frauenfelder, U. H. (2009). Behavioral and electrophysiological evidence for the impact of regional variation on phoneme perception. *Cognition*, 111(3):390–6.

Brunelliere, A. and Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and Language*, 125(1):82–93. 10.1016/j.bandl.2013.01.007.

Büchner, A., Schüssler, M., Battmer, R. D., Stöver, T., Lesinski-Schiedat, A., and Lenarz, T. (2009). Impact of low-frequency hearing. In *Audiology and Neurotology*, volume 14, pages 8–13. 10.1159/000206490.

Cabeza, R., Bruce, V., Kato, T., and Oda, M. (1999). The prototype effect in face recognition: Extension and limits. *Memory & Cognition*, 27(1):139–151.

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the mandarin triphthong /iau/. *Proceedings of the 8th ISSP*, pages 65–68.

Cai, S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2010). Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong

/iau/ and its pattern of generalization. *The Journal of the Acoustical Society of America*, 128(4):2033–48.

Cai, S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2011). Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *The Journal of Neuroscience*, 31(45):16483–16490.

Carbonell, K. M. and Lotto, A. J. (2014). Speech is not special... again. *Frontiers in Psychology*, 5:427. 10.3389/fpsyg.2014.00427.

Casserly, E. D. and Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdisciplinary Reviews. Cognitive Science*, 1(5):629–647.

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., and Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences of the United States of America*, 110(7):2653–8. 10.1073/pnas.1216827110.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, 13(11):1428–32. 10.1038/nn.2641.

Chase, R. A., Harvey, S., Standfast, S., Rapin, I., and Sutton, S. (1959). Comparison of the effects of delayed auditory feedback on speech and key tapping. *Science*, 129(3353):903–4. 10.1126/science.129.3353.903.

Chládková, K., Boersma, P., and Podlipský, V. J. (2009). On-line formant shifting as a function of F0. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, number 10, pages 464–467. International Speech Communication Association.

Chomsky, N. and Halle, M. (1968). *Sound pattern of English*. Harper & Row, New York.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–9.

Clopper, C. G., Pisoni, D. B., and Tierney, A. T. (2006). Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology*, 17(5):331–349.

Corney, D. and Lotto, R. B. (2007). What are lightness illusions and why do we see them? *PLoS Computational Biology*, 3(9):e180.

Creel, S. C. and Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, 5(5):190–204.

Cutler, A. and Stevens, J. R. (2006). Random forests for microarrays. *Methods in Enzymology*, 411:422–32.

Dahan, D., Drucker, S. J., and Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, 108(3):710–718.

D'Ausilio, A., Bufalari, I., Salmas, P., and Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, 48(7):882–887.

Davidson, P. R. and Wolpert, D. M. (2005). Widespread access to predictive models in the motor system: a short review. *Journal of Neural Engineering*, 2(3):S313.

Davis, M. H. and Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, 229(1-2):132–147.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and Mcgettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, 134(2):222–241.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*.

Devlin, J. T. and Watkins, K. E. (2007). Stimulating language: insights from TMS. *Brain*, 130(Pt 3):610–622. 10.1093/brain/awl331.

DeWitt, I. and Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8):E505–14. 10.1073/pnas.1113427109.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, 55:149–79.

Doupe, A. J. and Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience*, 22(1):567–631.

Doyle, J. (1987). Reliability of audiologists' ratings of the intelligibility of hearing-impaired children's speech. *Ear and Hearing*.

Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass*, 4(7):473–480.

Dunton, J., Bruce, C., and Newton, C. (2010). Investigating the impact of unfamiliar speaker accent on auditory comprehension in adults with aphasia. *International Journal of Language & Communication Disorders*.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. (1996). The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1393–1396. IEEE.

Eimas, P. D. and Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4(1):99–109.

Eisner, F. and McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, 67(2):224–38.

Eisner, F. and McQueen, J. M. (2006). Perceptual learning in speech: stability over time. *The Journal of the Acoustical Society of America*, 119(4):1950–3.

Elman, J. L. and McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27(2):143–165.

Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *The Journal of Neuroscience*, 33(26):10688–97.

Erickson, L. C., Zielinski, B. A., Zielinski, J. E., Liu, G., Turkeltaub, P. E., Leaver, A. M., and Rauschecker, J. P. (2014). Distinct cortical locations for integration of audiovisual speech and the mcgurk effect. *Frontiers in Psychology*, 5.

Evans, S. and Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cerebral Cortex*, 25(12):4772–4788.

Fahey, R. P., Diehl, R. L., and Traunmüller, H. (1996). Perception of back vowels: Effects of varying F1-F0 Bark distance. *The Journal of the Acoustical Society of America*, 99(4):2350.

Fant, G. (1960). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752–782.

Flagg, E. J., Oram Cardy, J. E., and Roberts, T. P. (2006). MEG detects neural consequences of anomalous nasalization in vowel-consonant pairs. *Neuroscience Letters*, 397(3):263–268.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, (14):3–28.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, 68(2):161–77.

Fowler, C. A. and Brown, J. M. (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception & Psychophysics*, 62(1):21–32.

Fowler, C. A. and Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4):742–754.

Franken, M. K., Hagoort, P., and Acheson, D. J. (2015). Modulations of the auditory M100 in an imitation task. *Brain and Language*, 142:18–23.

Galantucci, B., Fowler, C. A., and Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13(3):361–377.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1):110–125.

Gerrits, E. and Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, 66(3):363–376.

Ghyselinck, M., De Moor, W., and Brysbaert, M. (2000). Age-of-acquisition ratings for 2816 Dutch four- and five-letter nouns. *Psychologica Belgica*, 40(2):77–98.

Goldman, J. (2011). Easyalign: an automatic phonetic alignment tool under praat. In *Interspeech'11, 12th Annual Conference of the International Speech Communication Association*.

Goldstone, R. L. and Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78.

Goodale, M. A. and Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25. 10.1016/0166-2236(92)90344-8.

Greenwald, A. G. (1970). Sensory feedback mechanisms in performance control: with special reference to the ideo-motor mechanism. *Psychological Review*, 77(2):73–99.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species–29 years later. *The Journal of the Acoustical Society of America*, 87(6):2592–2605.

Grieser, D. and Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, 25(4):577–588.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27(3).

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, 72(1):43–53.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5):350–365.

Guenther, F. H. and Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, 25(5):408–422.

Halle, M. and Stevens, K. (1962). Speech recognition: A model and a program for research. *IRE Transactions on Information Theory*, 8(2):155–159.

Halle, M. and Stevens, K. N. (1959). Analysis by synthesis. In *Proc. Seminar on Speech Comprehension and Processing. Vol. 2*, pages 155–159.

Hannemann, R., Obleser, J., and Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research*, 1153(0):134–143. http://dx.note.org/10.1016/j.brainres.2007.03.069.

Hazan, V. and Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, 116(5):3108. 10.1121/1.1806826.

Heald, S. L. M. and Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, 8:35. 10.3389/fnsys.2014.00035.

Heinks-Maldonado, T. H., Mathalon, D. H., Gray, M., and Ford, J. M. (2005). Fine-tuning of auditory cortex during speech production. *Psychophysiology*, 42(2):180–190.

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, 21(7):1229–43. 10.1162/jocn.2009.21189.

Hickok, G. (2012a). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, 13(2):135–45. 10.1038/nrn3158.

Hickok, G. (2012b). The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders*, 45(6):393–402. 10.1016/j.jcomdis.2012.06.004.

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language and Cognitive Processes*, 29(1):2–20. 10.1080/01690965.2013.834370.

Hickok, G., Holt, L. L., and Lotto, A. J. (2009). Response to Wilson: What does motor cortex contribute to speech perception? *Trends in Cognitive Sciences*, 13(8):330–331.

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron*, 69(3):407–422. 10.1016/j.neuron.2011.01.019.

Hickok, G. and Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, 4(4):131–138.

Hickok, G. and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2):67–99.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402. 10.1038/nrn2113.

Hill, R. A. and Dunbar, R. I. (2003). Social network size in humans. *Human Nature*, 14(1):53–72.

Hirsh, I. J., Reynolds, E. G., and Joseph, M. (1954). Intelligibility of different speech materials. *The Journal of the Acoustical Society of America*, 26(4):530. 10.1121/1.1907370.

Hoemeke, K. A. and Diehl, R. L. (1994). Perception of vowel height: the role of F1-F0 distance. *The Journal of the Acoustical Society of America*, 96(2 Pt 1):661–74.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70.

Holst, E. and Mittelstaedt, H. (1950). Das reafferenzprinzip. *Naturwissenschaften*, 37(20):464–476.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, 16(4):305–12.

Holt, L. L. and Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science*, 17(1):42–46.

Holt, L. L., Lotto, A. J., and Otto, A. N. J. L. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, 72(5):1218–1227.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The Theory of Event Coding (TEC): a framework for perception and action planning. *Behavioral and Brain Sciences*, 24(5):849–878; discussion 878–937.

Houde, J. F. and Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, 279(5354):1213–1216.

Houde, J. F. and Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing research*, 45(2):295–310.

Houde, J. F. and Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, 5:82.

Howell, P. and Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication*, 10(2):163–169.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., and Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, 286:2526–8.

Ito, T., Coppola, J. H., and Ostry, D. J. (2016). Speech motor learning changes the neural response to both auditory and somatosensory signals. *Scientific Reports*, 6:25926.

Ito, T., Tiede, M., and Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences*, 106(4):1245–1248.

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, 59(4):434–446.

Jakobson, R., Fant, G., and Halle, M. (1963). *Preliminaries to speech analysis. The distinctive features and their correlates*. M.I.T. Press, Cambridge, MA, 3rd edition.

James, W. (1890). *The Principles of Psychology*. Harvard U. Press, Cambridge, MA, US.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage*, 14(1):S103–S109.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34(4):485–499.

Johnson, K., Strand, E. A., and D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. 10.1006/jpho.1999.0100.

Jones, J. A. and Munhall, K. G. (2005). Remapping auditory-motor representations in voice production. *Current Biology*, 15(19):1768–72.

Jordan, M. I. and Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16(3):307–354. 10.1016/0364-0213(92)90036-T.

Katseff, S., Houde, J., and Johnson, K. (2012). Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback? *Language and Speech*, 55(2):295–308.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9(6):718–27. 10.1016/S0959-4388(99)00028-8.

Kleinschmidt, D. F. and Jaeger, T. F. (2015a). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review*. 10.3758/s13423-015-0943-z.

Kleinschmidt, D. F. and Jaeger, T. F. (2015b). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2):148–203.

Knoblich, G. and Flach, R. (2001). Predicting the effects of actions: interactions of perception and action. *Psychological Science*, 12(6):467–472. 10.1111/1467-9280.00387.

Knoblich, G., Seigerschmidt, E., Flach, R., and Prinz, W. (2002). Authorship effects in the prediction of handwriting strokes: Evidence for action simulation during action perception. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 55(3):1027–1046.

Konishi, M. (1965). The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift für Tierpsychologie*, 22(7):770–783.

Kraljic, T. and Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51(2):141–78.

Kraljic, T. and Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, 13(2):262–8.

Kraljic, T. and Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56(1):1–15. 10.1016/j.jml.2006.07.010.

Kraljic, T., Samuel, A. G., and Brennan, S. E. (2008). First impressions and last resorts: how listeners adjust to speaker variability. *Psychological Science*, 19(4):332–338.

Krieger-Redwood, K., Gaskell, M. G., Lindsay, S., and Jefferies, E. (2013). The selective role of premotor cortex in speech perception: A contribution to phoneme judgements but not speech comprehension. *Journal of Cognitive Neuroscience*, 25(12):2179–2188.

Kuhl, P. K. (1989). Adults and infants show a "prototype effect" for speech sounds. *The Journal of the Acoustical Society of America*, 86(S1):S48.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2):93–107.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5(11):831–843.

Kunisaki, O. and Fujisaki, H. (1977). On the influence of context upon perception of voiceless fricative consonants. *Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo*, 11:85–91.

Laan, G. P. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22(1):43–65.

Lametti, D. R., Krol, S. A., Shiller, D. M., and Ostry, D. J. (2014a). Brief periods of auditory perceptual training can determine the sensory targets of speech motor learning. *Psychological Science*, page 0956797614529978.

Lametti, D. R., Nasir, S. M., and Ostry, D. J. (2012). Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback. *Journal of Neuroscience*, 32(27):9351–9358.

Lametti, D. R., Rochet-Capellan, a., Neufeld, E., Shiller, D. M., and Ostry, D. J. (2014b). Plasticity in the human speech motor system drives changes in speech perception. *Journal of Neuroscience*, 34:10339–10346.

Lane, H. and Webster, J. W. (1991). Speech deterioration in postlingually deafened adults. *The Journal of the Acoustical Society of America*, 89(2):859–866.

Latinus, M. and Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, 2(JUL):1–12.

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., and Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, 23(12):1075–1080.

Lawrence, M. A. (2011). ez: Easy analysis and visualization of factorial experiments. R package version 3.0-0.

Leonardo, A. and Konishi, M. (1999). Decrystallization of adult birdsong by perturbation of auditory feedback. *Nature*, 399(6735):466–70. 10.1038/20933.

Lev-Ari, S. (2015). How the size of our social network influences our semantic skills. *Cognitive Science*, pages 2050–2064. 10.1111/cogs.12317.

Levelt, W. J. (1989). *Speaking: From intention to articulation*. MIT Press Cambridge, MA.

Levelt, W. J., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(01):1–38.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104. 10.1016/0010-0277(83)90026-4.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Liberman, A., Delattre, P., and Cooper, F. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, 65(4):497–516.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431–461. 10.1037/h0020279.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Liberman, M. and Mattingly, G. (1989). A specialization for speech perception. *Science*, 243(4890):489–494.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., and Johansson, P. (2014). Speakers' acceptance of real-time speech exchange indicates that we use auditory feedback to specify the meaning of what we say. *Psychological Science*, 25(6):3–5.

Liu, R. and Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6):1783–1798.

Lively, S. E., Logan, J. S., and Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, 94(3):1242. 10.1121/1.408177.

Llinás, R. R. (2001). *I of the vortex: From neurons to self*. MIT press Cambridge, MA.

López, S., Riera, P., Assaneo, M. F., Eguía, M., Sigman, M., and Trevisan, M. a. (2013). Vocal caricatures reveal signatures of speaker identity. *Scientific Reports*, 3:3407. 10.1038/srep03407.

Lotto, A. J. (2000). Language acquisition as complex category formation. *Phonetica*, 57(2-4):189–96.

Lotto, A. J. and Kluender, K. R. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, 60(4):602–19.

Lotze, R. H. (1852). *Medicinische Psychologie oder Physiologie der Seele (Medical psychology or physiology of the soul)*. Weidmann'sche Buchhandlung, Leipzig.

Lubbers, M. and Torreira, F. (2013-2016). Praatalign: an interactive praat plug-in for performing phonetic forced alignment. https://github.com/dopefishh/praatalign. Version 1.9.

MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (2011). Probing the independence of formant control using altered auditory feedback. *The Journal of the Acoustical Society of America*, 129(2):955–65.

MacKain, K. S., Best, C. T., and Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, 2(4):229–250.

Magnuson, J. S. and Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, 33(2):391–409.

Mani, N. and Huettig, F. (2012). Prediction during language processing is a piece of cake-But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, 38(4):843–847. 10.1037/a0029284.

Mann, V. and Soli, S. D. (1991). Perceptual order and the effect of vocalic context of fricative perception. *Perception & Psychophysics*, 49(5):399–411.

Mann, V. A. and Repp, B. H. (1980). Influence of vocalic context on perception of the [š]-[s] distinction. *Perception & Psychophysics*, 28(3):213–228.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177–90.

Marques, L. M., Lapenta, O. M., Merabet, L. B., Bolognini, N., and Boggio, P. S. (2014). Tuning and disrupting the brain-modulating the mcgurk illusion with electrical stimulation. *Frontiers in Human Neuroscience*, 8.

Massaro, D. W. and Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, 2(1):15–35. 10.1016/0167-6393(83)90061-4.

Mathworks (2012). MATLAB and Statistics Toolbox.

Mattar, A. A. G., Nasir, S. M., Darainy, M., and Ostry, D. J. (2011). *Enhancing Performance for Action and Perception - Multisensory Integration, Neuroplasticity and Neuroprosthetics, Part I*, volume 191 of *Progress in Brain Research*. Elsevier.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1–86. 10.1016/0010-0285(86)90015-0.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech*, page 0023830914565191.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.

McHenry, M. (2011). An exploration of listener variability in intelligibility judgments. *American Journal of Speech-Language Pathology*, 20(2):119–123.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, 17(2):433–43.

McQueen, J. M. (2005). Spoken-word recognition and production: regular but not inseparable bedfellows. In *Twenty-first century psycholinguistics: Four cornerstones*, pages 229–244. Lawrence Erlbaum Associates.

McQueen, J. M., Cutler, A., and Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, 30(6):1113–26.

Menzerath, P. and de Lacerda, A. (1933). *Koartikulation, Steuerung und Lautabgrenzung*. Ferd. Dümmlers Verlag, Berlin and Bonn.

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders*, 48(6):601–612. 10.1111/1460-6984.12061.

Mitsuya, T., MacDonald, E. N., Munhall, K. G., and Purcell, D. W. (2015). Formant compensation for auditory feedback with English vowels. *The Journal of the Acoustical Society of America*, 138(1):413–24. 10.1121/1.4923154.

Mitterer, H. and Blomert, L. (2003). Coping with phonological assimilation in speech perception: evidence for early compensation. *Perception & Psychophysics*, 65(6):956–69.

Moineau, S., Dronkers, N. F., and Bates, E. (2005). Exploring the processing continuum of single-word comprehension in aphasia. *Journal of Speech, Language, and Hearing Research*, 48(4):884–896.

Morey, R. D., Rouder, J. N., and Jamil, T. (2015). BayesFactor: Computation of Bayes Factors for common designs. R package version 0.9.12-2.

Mostert, P., Kok, P., and de Lange, F. P. (2015). Dissociating sensory from decision processes in human perceptual decision making. *Scientific Reports*, 5.

Moulin-Frier, C. and Arbib, M. A. (2013). Recognizing speech in a novel accent: the motor theory of speech perception reframed. *Biological Cybernetics*, 107(4):421–447. 10.1007/s00422-013-0557-3.

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J. L., and Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor, and perceptuo-motor theories of speech perception: An exploratory Bayesian modelling study. *Language and Cognitive Processes*, 27(7-8):1240–1263. 10.1080/01690965.2011.645313.

Nasir, S. M. and Ostry, D. J. (2009). Auditory plasticity and speech motor learning. *Proceedings of the National Academy of Sciences of the United States of America*, 106(48):20470–5.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, 85:2088–2113. 10.1121/1.397861.

Netwerk Naamkunde (2015). Available at `http://www.naamkunde.net/`.

Nguyen, N., Dufour, S., and Brunellière, A. (2012). Does imitation facilitate word recognition in a non-native regional accent? *Frontiers in Psychology*, 3:480.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, 18(1):62–85. 10.1177/0261927X99018001005.

Niziolek, C. A. and Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *The Journal of Neuroscience*, 33(29):12090–8.

Niziolek, C. A., Nagarajan, S. S., and Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *The Journal of Neuroscience*, 33(41):16110–6.

Norris, D. and McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357–95.

Norris, D., McQueen, J. M., and Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(03):299–325.

Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2):204–38.

Nourouzpour, N., Salomonczyk, D., Cressman, E. K., and Henriques, D. Y. P. (2015). Retention of proprioceptive recalibration following visuomotor adaptation. *Experimental Brain Research*, 233(3):1019–29.

Nozari, N., Dell, G. S., and Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology*, 63(1):1–33. 10.1016/j.cogpsych.2011.05.001.

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., and Wagenmakers, E.-J. (2015). BayesMed: Default Bayesian hypothesis tests for correlation, partial correlation, and mediation. R package version 1.0.1.

Nusbaum, H. and Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In Johnson, K. and Mullennix, J., editors, *Talker Variability in Speech Processing*, pages 109–132. Academic Press, San Diego. 10.1121/1.2028337.

Nuttall, H. E., Kennedy-Higgins, D., Hogan, J., Devlin, J. T., and Adank, P. (2016). The effect of speech distortion on the excitability of articulatory motor cortex. *NeuroImage*, 128:218–226.

Nygaard, L. C. and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3):355–376.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5(1):42–46.

Obleser, J. and Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, 13(1):14–19.

Obleser, J., Elbert, T., Lahiri, A., and Eulitz, C. (2003a). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cognitive Brain Research*, 15(3):207–213.

Obleser, J., Lahiri, A., and Eulitz, C. (2003b). Auditory-evoked magnetic field codes place of articulation in timing and topography around 100 milliseconds post syllable onset. *NeuroImage*, 20(3):1839–1847.

Obleser, J., Lahiri, A., and Eulitz, C. (2004). Magnetic Brain Response Mirrors Extraction of Phonological Features from Spoken Vowels. *Journal of Cognitive Neuroscience*, 16(1):31–39.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: Open source software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological data. *Computational Intelligence and Neuroscience*.

O'Regan, J. K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(05):939–973.

Ostry, D. J., Darainy, M., Mattar, A. A. G., Wong, J., and Gribble, P. L. (2010). Somatosensory plasticity and motor learning. *The Journal of Neuroscience*, 30(15):5384–93.

Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L. E., and Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature*, 392(6678):811–814. 10.1038/33918.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, 4:559.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., and Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, 69(3):183–195.

Paulus, M., Hunnius, S., van Elk, M., and Bekkering, H. (2012). How learning to shake a rattle affects 8-month-old infants' perception of the rattle's sound: electrophysiological evidence for action-effect binding in infancy. *Developmental Cognitive Neuroscience*, 2(1):90–6.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., and Zandipour, M. (2004a). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America*, 116(4):2338–2344.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., and Guenther, F. H. (2004b). The distinctness of speakers' /s/-/sh/ contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language, and Hearing Research*, 47(6):1259–1269.

Pickering, M. J. and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3):105–110. 10.1016/j.tics.2006.12.002.

Pickering, M. J. and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(04):329–347.

Picton, T. (2013). Hearing in time. *Ear and Hearing*, 34(4):385–401.

Pinget, A.-F., Bosker, H. R., Quene, H., and de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, 31(3):349–365. 10.1177/0265532214526177.

Plant, G. and Hammarberg, B. (1983). Acoustic and perceptual analysis of the speech of the deafened. *Speech Transmission Laboratory, Quarterly Progress and Status Report (Stockholm)*, 2(3):85–107.

Poeppel, D., Idsardi, W. J., and Van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493):1071–1086.

Poeppel, D. and Monahan, P. J. (2008). Speech perception: Cognitive foundations and cortical implementation. *Current Directions in Psychological Science*, 17(2):80–85. 10.1111/j.1467-8721.2008.00553.x.

Poeppel, D. and Monahan, P. J. (2011). Feedforward and feedback in speech perception: Revisiting analysis by synthesis. *Language and Cognitive Processes*, 26(7):935–951. 10.1080/01690965.2010.493301.

Poeppel, D., Phillips, C., Yellin, E., Rowley, H. A., Roberts, T. P., and Marantz, A. (1997). Processing of vowels in supratemporal auditory cortex. *Neuroscience Letters*, 221(2-3):145–148.

Potter, R. K. (1946). Introduction to technical discussions of sound portrayal. *The Journal of the Acoustical Society of America*, 18(1):1–3.

Prinz, W. (1990). A common coding approach to perception and action. In Neumann, O. and Prinz, W., editors, *Relationships Between Perception and Action*, pages 167–201. Springer Berlin Heidelberg.

Purcell, D. W. and Munhall, K. G. (2006). Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America*, 120(2):966–77.

Puria, S. and Rosowski, J. J. (2012). Békésy's contributions to our present understanding of sound conduction to the inner ear. *Hearing Research*, 293(1-2):21–30.

Quatieri, T. and McAulay, R. J. (1986). Speech transformations based on a sinusoidal representation. *IEEE Transactions (ASSP)*, ASSP-34(6):1449–1464.

R Development Core Team (2013). R Development Core Team.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, 8(4):516–521.

Rauschecker, J. P. and Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6):718–24. 10.1038/nn.2331.

Reinisch, E., Wozny, D. R., Mitterer, H., and Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of phonetics*, 45:91–105.

Remez, R. E., Dubowski, K. R., Broder, R. S., Davids, M. L., Grossman, Y. S., Moskalenko, M., Pardo, J. S., and Hasbun, S. M. (2011). Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3):968.

Repp, B. H. and Mann, V. A. (1978). Influence of vocalic context on perception of the [s]-[š] distinction. *The Journal of the Acoustical Society of America*, 64(S1):S17–S17.

Repp, B. H. and Mann, V. A. (1981). Perceptual assessment of fricative–stop coarticulation. *The Journal of the Acoustical Society of America*, 69(4):1154–63.

Roberts, S. G., Torreira, F., and Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, 6:509.

Rochet-Capellan, A. and Ostry, D. J. (2011). Simultaneous acquisition of multiple auditory–motor transformations in speech. *Journal of Neuroscience*, 31(7):2657–2662.

Rochet-Capellan, A., Richer, L., and Ostry, D. J. (2012). Nonhomogeneous transfer reveals specificity in speech motor learning. *Journal of Neurophysiology*, 107(6):1711–7.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, 4(3):328–350.

Rosch, E. and Lloyd, B. B. (1978). Cognition and Categorization. *Lloydia Cincinnati*, pp:27–48.

Ross, B., Jamali, S., and Tremblay, K. L. (2013). Plasticity in neuromagnetic cortical responses suggests enhanced auditory object representation. *BMC Neuroscience*, 14(1):151.

Rubin, D. L. and Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative english-speaking teaching assistants. *International Journal of Intercultural Relations*, 14(3):337–353.

Runeson, S. and Frykholm, G. (1983). Kinematic specification of dynamics as an informational basis for person-and-action perception: Expectation, gender recognition, and deceptive intention. *Journal of Experimental Psychology: General*, 112(4):585–615.

Sagan, C., Druyan, A., and Soter, S. (1980). The backbone of night. [Television series episode]. In Cosmos: A Personal Voyage. PBS.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, 31(4):307–314.

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology*, 18(4):452–499.

Samuel, A. G. and Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics*, 71(6):1207–18.

Sancier, M. L. and Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, 25(4):421–436.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., and Weiller,

C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(46):18035–40. 10.1073/pnas.0805234105.

Savariaux, C., Perrier, P., Orliaguet, J. P., and Schwartz, J. L. (1999). Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis. *The Journal of the Acoustical Society of America*, 106(1):381–93.

Schouten, B., Gerrits, E., and van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, 41(1):71–80.

Schuerman, W. L., Meyer, A., and McQueen, J. M. (2015). Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker. *PLoS ONE*, 10(7):1–18.

Schütz-Bosbach, S. and Prinz, W. (2007). Perceptual resonance: action-induced modulation of perception. *Trends in Cognitive Sciences*, 11(8):349–355. 10.1016/j.tics.2007.06.005.

Schwartz, J.-L. L., Basirat, A., Ménard, L., and Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5):336–354. 10.1016/j.jneuroling.2009.12.004.

Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123(12).

Scott, S. K., McGettigan, C., and Eisner, F. (2009). A little more conversation, a little less action - candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience*, 10(4):295–302. 10.1038/nrn2603.

Scott, S. K., McGettigan, C., and Eisner, F. (2013). The neural basis of links and dissociations between speech perception and production. In Bolhuis, J. J. and Everaert, M., editors, *Birdsong, Speech and Language: Exploring the Evolution of Mind and Brain*, pages 277–295. MIT Press.

Sesterhenn, G. and Breuninger, H. (1978). On the influence of the middle ear muscles upon changes in sound transmission. *Archives of Oto-Rhino-Laryngology*, 221(1):47–60.

Shankweiler, D. and Fowler, C. A. (2015). Seeking a reading machine for the blind and discovering the speech code. *History of Psychology*, 18(1):78–99. 10.1037/a0038299.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304. 10.1126/science.270.5234.303.

Sheehan, K. A., McArthur, G. M., and Bishop, D. V. (2005). Is discrimination training necessary to cause changes in the P2 auditory event-related brain potential to speech sounds? *Cognitive Brain Research*, 25(2):547–553.

Sheldon, A. (1985). The relationship between production and perception of the /r/-/l/ contrast in Korean adults learning English: A reply to Borden, Gerber, and Milsark. *Language Learning*, 35(1):107–113. 10.1111/j.1467-1770.1985.tb01018.x.

Sheldon, A. and Strange, W. (1982). The acquisition of /r/ and /I/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, 3:243–261. 10.1017/S0142716400001417.

Shestakova, A., Brattico, E., Soloviev, A., Klucharev, V., and Huotilainen, M. (2004). Orderly cortical representation of vowel categories presented by multiple exemplars. *Brain Research*, 21(3):342–50.

Shiller, D. M., Sato, M., Gracco, V. L., and Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *The Journal of the Acoustical Society of America*, 125(February):1103–1113.

Shin, Y. K., Proctor, R. W., and Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychological Bulletin*, 136(6):943–974.

Sitek, K. R., Mathalon, D. H., Roach, B. J., Houde, J. F., Niziolek, C. A., and Ford, J. M. (2013). Auditory cortex processes variation in our own speech. *PLoS ONE*, 8(12):e82925.

Skipper, J. I., Devlin, J. T., and Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language*, 164:77–105.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience*, 32(25):8443–8453. 10.1523/jneurosci.5069-11.2012.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (2014). Top-down influences of written text on perceived clarity of degraded speech. *Journal of Experimental Psychology: Human Perception and Performance*, 40(1):186–99. 10.1037/a0033206.

Sonderegger, M. and Yu, A. (2010). A rational account of perceptual compensation for coarticulation. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.*, pages 275–280. Cognitive Science Society.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, 43(6):482.

Stasenko, A., Bonn, C., Teghipco, A., Garcea, F. E., Sweet, C., Dombovy, M., McDonough, J., and Mahon, B. Z. (2015). A causal test of the motor theory of speech perception: a case of impaired speech production and spared speech perception. *Cognitive Neuropsychology*, 32(2):38–57. 10.1080/02643294.2015.1035702.

Stasenko, A., Garcea, F. E., and Mahon, B. Z. (2013). What happens to the motor theory of perception when the motor system is damaged? *Language and Cognition*, 5(2-3):225–238.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4):1872–91.

Stevens, K. N. and Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In Walthen-Dunn, W., editor, *Models for the perception of speech and visual form*, pages 88–102. MIT Press Cambridge, MA.

Stevens, S. S. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185.

Strand, E. A. and Johnson, K. (1996). Gradient and Visual Speaker Normalization in the Perception of Fricatives. In Gibbon, D., editor, *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielfelt*, pages 14–26, Berlin. Mouton.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25.

Strobl, C., Hothorn, T., and Zeileis, A. (2009). Party on! A new, conditional variable importance measure for random forests available in the party package. *The R Journal*, 1/2:14–17.

Sumner, M. (2015). The social weight of spoken words. *Trends in Cognitive Sciences*, 19(5):238–239.

Sumner, M., Kim, S., King, E., and McGowan, K. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Frontiers in Psychology*, 4:1015.

Tagliamonte, S. A. and Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(02):135–178.

Tavabi, K., Obleser, J., Dobel, C., and Pantev, C. (2007). Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing. *European Journal of Neuroscience*, 25(10):3155–3162.

Tian, X. and Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, 1:166.

Tian, X. and Poeppel, D. (2013). The effect of imagination on stimulation: The functional specificity of efference copies in speech processing. *Journal of Cognitive Neuroscience*, 25(7):1020–1036.

Tourville, J. A., Cai, S., and Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech. In *Proceedings of Meetings on Acoustics*, volume 19, page 060180. Acoustical Society of America.

Tourville, J. A. and Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, 26(7):952–981. 10.1080/01690960903498424.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *The Journal of the Acoustical Society of America*, 69(5):1465.

Tremblay, K., Kraus, N., Carrell, T. D., and McGee, T. (1997). Central auditory system plasticity: generalization to novel stimuli following listening training. *The Journal of the Acoustical Society of America*, 102(6):3762–73.

Tremblay, K., Kraus, N., McGee, T., Ponton, C., and Otis, B. (2001). Central auditory plasticity: changes in the N1-P2 complex after speech-sound training. *Ear and Hearing*, 22(2):79–90.

Tremblay, K. L., Inoue, K., McClannahan, K., and Ross, B. (2010). Repeated stimulus exposure alters the way sound is encoded in the human brain. *PLoS ONE*, 5(4):e10283.

Tremblay, K. L., Ross, B., Inoue, K., McClannahan, K., and Collet, G. (2014). Is the auditory evoked P2 response a biomarker of learning? *Frontiers in Systems Neuroscience*, 8:28.

Tremblay, K. L., Shahin, A. J., Picton, T., and Ross, B. (2009). Auditory training alters the physiological detection of stimulus-specific cues in humans. *Clinical Neurophysiology*, 120(1):128–135.

Tschida, K. and Mooney, R. (2012). The role of auditory feedback in vocal learning and maintenance. *Current Opinion in Neurobiology*, 22(2):320–7.

Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., and Henry, B. A. (2004). Speech recognition in noise for cochlear implant listeners: benefits of residual acoustic hearing. *The Journal of the Acoustical Society of America*, 115:1729–1735. 10.1121/1.1687425.

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., and Sommers, M. S. (2013). Reading your own lips: common-coding theory and visual speech perception. *Psychonomic Bulletin & Review*, 20(1):115–9. 10.3758/s13423-012-0328-5.

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., and Sommers, M. S. (2015). The self-advantage in visual speech processing enhances audiovisual speech recognition in noise. *Psychonomic Bulletin & Review*, 22(4):1048–53.

Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., and Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, 20(4):580–591.

van Linden, S., Stekelenburg, J. J., Tuomainen, J., and Vroomen, J. (2007). Lexical effects on auditory speech perception: An electrophysiological study. *Neuroscience Letters*, 420(1):49–52.

van Linden, S. and Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, 33(6):1483.

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, 122(4):2306. 10.1121/1.2773966.

Viswanathan, N., Fowler, C. A., and Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review*, 16(1):74–9.

Viswanathan, N., Magnuson, J. S., and Fowler, C. A. (2010). Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance*, 36(4):1005–15.

von Helmholtz, H. (1878/1977). The facts in perception. In *Epistemological Writings*, pages 115–185. Springer.

Vongphoe, M. and Zeng, F.-G. (2005). Speaker recognition with temporal cues in acoustic and electric hearing. *The Journal of the Acoustical Society of America*, 118:1055–1061. 10.1121/1.1944507.

Vurma, A. (2014). The timbre of the voice as perceived by the singer him-/herself. *Logopedics Phoniatrics Vocology*, 39(1):1–10.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. a. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6):627–633. 10.1177/1745691612463078.

Watkins, K. E., Strafella, A. P., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8):989–994.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63.

Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the english [s]–[š] boundary. *The Journal of the Acoustical Society of America*, 69(1):275–282.

Wilson, M. and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131(3):460–473.

Wilson, S. M. (2009). Speech perception when the motor system is compromised. *Trends in Cognitive Sciences*, 13(8):329–330.

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences*, 1(6):209–216. 10.1016/S1364-6613(97)01070-X.

Wolpert, D. M. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3(Supp):1212–1217. 10.1038/81497.

Xu, Y. (2007). FormantPro.praat. Retrieved from `http://www.phon.ucl.ac.uk/home/yi/FormantPro/`.

Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin*, 60(3):213–232. 10.1037/h0044155.

Yi, H.-G., Phelps, J. E., Smiljanic, R., and Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America*, 134(5):EL387–EL393.

Zhang, T., Dorman, M. F., and Spahr, A. J. (2010). Information from the voice fundamental frequency (F0) region accounts for the majority of the benefit when acoustic stimulation is added to electric stimulation. *Ear and Hearing*, 31:63–69. 10.1097/AUD.0b013e3181b7190c.

Zhu, Q. and Bingham, G. P. (2014). Seeing where the stone is thrown by observing a point-light thrower: Perceiving the effect of action is enabled by information, not motor experience. *Ecological Psychology*, 26(4):229–261.

Nederlandse samenvatting

Spreken is articulatorisch gezien net acrobatiek en levert akoestisch een zeer complex signaal op dat luisteraars met verbluffend gemak weten te ontcijferen. Het is vaak geopperd dat luisteraars deze uitzonderlijke prestatie kunnen leveren dankzij hun eigen ervaring met het produceren van spraak. Als we onze kennis van hoe we zelf spreken gebruiken om te begrijpen wat anderen zeggen, beïnvloedt onze eigen spraak misschien ook wel hoe we de spraak van anderen waarnemen. Mijn onderzoek heeft aangetoond dat sprekers die op elkaar lijken, elkaar beter kunnen verstaan. Echter, de experimenten laten ook zien dat we een gemiddelde spreker nog beter verstaan dan een spreker die op onszelf lijkt. Dit wijst erop dat luisteraars putten uit hun auditieve ervaring met verschillende sprekers in de gemeenschap. Deze ogenschijnlijk tegenstrijdige resultaten komen overeen met recente neurobiologische modellen van spraakperceptie, die stellen dat verschillende spraakfuncties uiteenlopende hersennetwerken aanspreken.

English summary

When we speak, incredible articulatory acrobatics create complex acoustic signals that listeners decode with astonishing ease. Many have suggested that humans accomplish this feat by drawing upon experience acquired during speech production to accomplish speech perception, i.e., we use our knowledge of how we speak to understand what others are saying. If so, then the way that we speak may influence how we perceive others' speech. My research revealed that the way we produce sounds may influence how we perceive sounds produced by others, and that similar talkers are more intelligible to each other. Yet the experiments also revealed that an average talker is more understandable than a similar talker, suggesting that listeners draw on their auditory experience with different talkers in their community. These seemingly conflicting results concur with recent neurobiological models of speech perception that propose that different speech tasks recruit divergent networks in the brain.

CV

William Lawrence Schuerman (born in Sonoma on September 28th, 1984) received his bachelor's degree in linguistics from the University of California, Berkeley, then traveled to South Korea and Japan as an English teacher. He began to specialize in phonetics and speech perception while obtaining his master's degree from Utrecht University.

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda I. van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed class distinction in spoken-word recognition. *Alette Petra Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk J.Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie H. van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*

15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene KrÃď mer*

16. Language-specific listening: The case of phonetic sequences. *Andrea Christine Weber*

17. Moving eyes and naming objects. *Femke Frederike van der Meulen*

18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*

19. Morphology in speech comprehension. *Kerstin Mauth*

20. Morphological families in the mental lexicon. *Nivja Helena de Jong*

21. Fixed expressions and the production of idioms. *Simone Annegret Sprenger*

22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria. *Birgit Hellwig*

23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*

24. Contextual influences on spoken-word processing: An electrophysiological approach. *Danielle van den Brink*

25. Perceptual relevance of prevoicing in Dutch. *Petra Martine van Alphen*

26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*

27. Producing complex spoken numerals for time and space. *Marjolein HenriÃńtte Wilhelmina Meeuwissen*

28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *RachÃĺl Jenny Judith Karin Kemps*

29. At the same timeâĂę: The expression of simultaneity in learner varieties. *Barbara SchmiedtovÃą*

30. A grammar of Jalonke argument structure. *Friederike Lüpke*

31. Agrammatic comprehension: An electrophysiological approach. *Marijtje Elizabeth Debora Wassenaar*

32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*

33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*

34. Phonetic and lexical processing in a second language. *Mirjam Elisabeth Broersma*

35. Retrieving semantic and syntactic word properties: ERP studies on the time course in language comprehension. *Oliver Müller*

36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*

37. Sensitivity to detailed acoustic information in word recognition. *Keren Batya Shatzman*

38. The relationship between spoken word production and comprehension. *Rebecca ÃŰzdemir*

39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*

40. The acquisition of phonological structure: Distinguishing contrastive from non-constrative variation. *Christiane Dietrich*

41. Cognitive cladistics and the relativity of spatial cognition. *Daniel Haun*

42. The acquisition of auditory categories. *Martijn Bastiaan Goudbeek*

43. Affix reduction in spoken Dutch: Probabilistic effects in production and perception. *Mark Plumaekers*

44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Madalla Kooijman*

45. Space and iconicity in German sign language (DGS). *Pamela M. Perniss*

46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*

47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*

48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*

49. Phoneme inventories and patterns of speech sound perception. *Anita Eva Wagner*

50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*

51. A grammar of Savosavo: A Papuan language of the Solomon Islands. *Claudia Ursula Wegener*

52. Prosodic structure in speech production and perception. *Claudia Kuzla*

53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*

54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*

55. Processing the fine temporal structure of spoken words. *Eva Reinisch*

56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*

57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*

58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*

59. Lexical interactions in non-native speech comprehension: Evidence from electroencephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*

60. Processing casual speech in native and non-native language. *Annelie Tuinman*

61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Payton Robinson*

62. Evidentiality and intersubjectivity in YurakarÃł: An interactional account. *Sonja Gipper*

63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*

64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*

65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*

66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*

67. The relevance of early word recognition: Insights from the infant brain. *Caroline Mary Magteld Junge*

68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias Johannes Sjerps*

69. Structuring language: Contributions to the neurocognition of syntax. *Katrien Rachel Segaert*

70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*

71. Gaze behavior in face-to-face interaction. *Federico Rossano*

72. Sign-spatiality in Kata Kolok: How a village sign language of Bali inscribes its signing space. *Connie de Vos*

73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*

74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*

75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*

76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*

77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*

78. Contributions of executive control to individual differences in word production. *Zeshu Shao*

79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*

80. High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*

81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*

82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*

83. The many ways listeners adapt to reductions in casual speech. *Katja Pöllmann*

84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*

85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*

86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*

87. The acquisition of morphophonological alternations across languages. *Helen Buckler*

88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*

89. Rediscovering a forgotten language. *Jiyoun Choi*

90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*

91. Infants'understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*

92. Information structure in Avatime. *Saskia van Putten*

93. Switch reference in Whitesands. *Jeremy Hammond*

94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*

95. Acquisition of spatial language by signing and speaking children: A comparison of Turkish sign language (TID) and Turkish. *Beyza Sumer*

96. An ear for pitch: On the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*

97. lncrementality and Flexibility in Sentence Production. *Maartje van de Velde*

98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*

99. The request system in Italian interaction. *Giovanni Rossi*

100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*

101. Assessing birth language memory in young adoptees. *Wencui Zhou*

102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*

103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*

104. Conversation electrified: The electrophysiology of spoken speech act recognition. *Rósa Signý Gisladottir*

105. Modelling multimodal language processing. *Alastair Charles Smith*

106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*

107. Switch reference in Whitesands. *Jeremy Hammond*

108. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*

109. An ear for pitch: On the effects of experience and aptitude in processing pitch in language and music. *Salomi S. Asaridou*

110. Semantic specificity of perception verbs in Maniq. *Ewelina Wnuk*

111. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntactic, and orthographic effects. *Malte Viebahn*

112. Situational variation in non-native communication: Studies into register variation, discourse management and pronunciation in Spanish English. *Huib Kouwenhoven*

113. Sustained attention in language production. *Suzanne R. Jongman*

114. Events in language and thought: The case of serial verb constructions in Avatime. *Rebecca Defina*

115. Deciphering common and rare genetic effects on reading ability. *Amaia Carrión Castillo*

116. Nativeness, dominance, and the flexibility of listening to spoken language. *Laurence Bruggeman*

117. On the identification of FOXP2 gene enhancers and their role in brain development. *Martin Becker*

118. Comprehending comprehension: Insights from neuronal oscillations on the neuronal basis of language. *Nietzsche H. L. Lam*

119. Music and language comprehension in the brain. *Richard Kunert*

120. The biology of variation in anatomical brain asymmetries. *Tulio Guadalupe*